Analyses génomiques de données sur le vieillissement cutané

Vincent Laville

le cnam

ÉCOLE DOCTORALE Arts et Métiers

Laboratoire Génomique Bioinformatique et Applications (GBA)

THÈSE présentée par :

Vincent LAVILLE

soutenue le : **30 janvier 2015**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline/ Spécialité : Bioinformatique

Analyses génomiques de données sur le vieillissement cutané

THÈSE dirigée par :

M. ZAGURY Jean-François	Professeur, CNAM, Paris

RAPPORTEURS :

Mme BONNEFOND Amélie	Docteur, Institut Pasteur, CNRS ,UMR8199, Lille
M. EZZEDINE Khaled	Professeur, Hôpital St-André, Bordeaux

JURY :

Mme GUINOT Christiane	Docteur, Laboratoires Servier
Mme LATREILLE Julie	Docteur, Société Chanel R&T
M. THERWATH Amu	Professeur, Université Paris-Diderot

« La science ne sert guère qu'à nous donner une idée de l'étendue de notre ignorance. »

Félicité de Lamennais

Remerciements

Je tiens tout d'abord à remercier vivement le Professeur Jean-François Zagury, pour m'avoir permis de réaliser ce travail de thèse, pour sa confiance et sa disponibilité tout au long de ces années.

Je remercie aussi le Dr Amélie Bonnefond et le Dr Khaled Ezzedine d'avoir accepté de juger mon travail et d'être rapporteur de cette thèse.

Je tiens également à remercier le Dr Christiane Guinot, le Dr Julie Latreille et le Pr Amu Therwath d'avoir accepté de participer au jury de cette thèse.

Je souhaite également exprimer ma gratitude à l'équipe R&T de Chanel , Dr Frédérique Morizot, Dr Julie Latreille, Dr Randa Jdid, Dr Gaëlle Gendronneau et Dr Irina Berlin, avec qui j'ai collaboré pour réaliser ce travail, et qui m'ont fait confiance pour l'analyse des données et ont toujours répondu avec beaucoup de gentillesse et de disponibilité.

J'aimerais également remercier le laboratoire Génomique Bioinformatique et Applications du CNAM pour les bons moments passés ensemble. En particulier, Christiane, Janine, Jean-Louis et Juanjo pour leur gentillesse et leur disponibilté.

Dans un premier temps, l'équipe Génomique, par ordre alphabétique : Cedcoul pour nos désaccords « parisianno-sud-américains », Damien pour m'avoir fait (presque !) aimé la country, Hervé pour sa formation pointue au « trollage », Josselin pour sa bonne humeur et sa créativité artistique, Lieng pour sa découverte d'univers orientaux particuliers, Marc et sa chance du débutant, Olivier pour ses innombrables citations, Pierre pour son habileté musicale, Sigrid pour sa sérénité à toutes épreuves et ses bons mots affectifs, Sophie pour le mythe qu'elle incarne au sein du laboratoire, Toufik pour m'avoir fait découvrir « Conardo ».

Vient ensuite l'équipe Drug Design : Charly pour son bon esprit, Héléne pour nos querelles perpétuelles, Matthieu pour avoir mis un peu de métal dans ce monde de brutes, Nathalie pour son éternelle bonne humeur et son soutien inconditionnel, Nesrine pour la découverte de la gastronomie tunisienne.

Enfin, l'équipe « Cytokine » : Gaby pour ses nombreuses présentations mémorables, Julie pour ses positions super-héroïques, Hadley pour son entrain inconditionnel, Lucille pour nos combats divinatoires footballistiques.

Je n’oublie pas les stagiaires passés par le laboratoire qui ont tous apporté un vent de fraîcheur et d’innocence dans l’équipe.

Je remercie également à remercier mes amis qui m’ont toujours supporté et les bons moments passés ensemble. A celui que nous n’oublions pas. Au « mollah ».

Enfin, même si ce n’est pas assez, un grand merci à ma famille qui m’a toujours soutenu et m’a permis d’être la personne que je suis. J’espère les rendre fiers.

Résumé

La compréhension des mécanismes moléculaires du vieillissement est une préoccupation majeure de la communauté scientifique notamment pour son intérêt en terme de santé publique. Le vieillissement cutané constitue un excellent modèle d'étude de ces mécanismes. En plus de facteurs environnementaux fortement influents, les facteurs génétiques y jouent un rôle fondamental. Les avancées technologiques en biologie moléculaire et en bioinformatique ont rendu possible le développement d'étude d'association « génome-entier » qui permettent d'identifier des associations statistiques entre les variants génétiques sur tout le génome et des phénotypes particuliers du vieillissement cutané. Ces associations, en amenant à une meilleure compréhension des mécanismes moléculaires de vieillissement, devraient permettre de développer de nouvellees stratégies d'intervention sur le vieillissement cutané et général.

Au cours de ma thèse, j'ai eu accès à une cohorte exceptionnelle de 502 femmes caucasiennes dont les caractéristiques du visage (rides, relâchement, lentigines, photo-vieillissement) ont été particulièrement bien définies. J'ai pu réaliser deux études d'association « génome-entier » pour identifier des polymorphismes associés à la sévérité des lentigines sur le visage d'une part et de l'affaissement de la paupière supérieure d'autre part. La première a mis en évidence le rôle du système immunitaire, et en particulier du gène *HLA-C*, dans la sévérité des lentigines. La seconde a identifié deux associations significatives avec la sévérité de l'affaissement de la paupière dont une avec des SNPs situés dans un intron du gène *H2AFY2*, codant pour une histone. Afin de compléter ces analyses, j'ai aussi recherché les voies de signalisation biologiques significativement associées aux caractéristiques du vieillissement cutané. Cette analyse a entre autre permis de souligner le rôle de la mélanogenèse et des mécanismes de réparation de l'ADN. Ce travail doit être répliqué par l'intermédiaire d'autres méthodes.

Ces résultats, en cours de publication dans des revues scientifiques internationales à comité de lecture, ouvrent de nouvelles perspectives dans la compréhension des mécanismes inhérents au vieillissement cutané et plus généralement, au vieillissement.

Mots clés : étude d'association « génome-entier », GWAS, SNP, voie de signalisation biologique, vieillissement cutané

Résumé en anglais

A better understanding of the molecular mechanisms responsible for global ageing is a major issue for the scientific community especially for its impact on public health. Skin ageing is an excellent model to study such mechanisms. In addition to the environmental factors which have a strong influence, genetic factors play a key role in these mechanisms. Technological progress in molecular biology and bioinformatics have made possible the development of genome-wide association studies which aim to identify statistical associations between genetic polymorphisms and skin ageing phenotypes. These associations, by leading to a better understanding of ageing molecular mechanisms, should allow the development of new interventional strategies for skin and global ageing.

During my PhD, I have had access to an exceptional cohort of 502 Caucasian women with very well-defined facial characteristics: wrinkles, sagging, lentigines, photo-ageing. I performed two genome-wide associations studies on the severity of lentigines and of the superior eyelid sagging. The first study pointed to a role of the immune system, and especially the *HLA-C* gene, in the severity of solar lentigines on the face. The second one identified two significant associations with the severity of superior eyelid drooping, one with SNPs located in an intronic part of the *H2AFY2* gene, encoding for a histone. To complete these analyses, I also looked for biological pathways significantly associated with the skin ageing characteristics. This latest analysis underlined the role played by melanogenesis and by mechanisms of DNA repair. This work must be replicated with other methods.

These results, under publications in peer-review international scientific journals, open new insights into the understanding of molecular mechanisms of skin ageing and of global ageing.

Keywords: genome-wide association study, GWAS, SNP, biological pathway, skin ageing

Table des matières

Liste des tableaux

Liste des figures

Liste des abbréviations

ACP : Analyse en Composantes Principales

ADN : Acide DésoxyriboNucléique

ARNm : Acide RiboNucléique Messager

bp : paire de base

CNV : Copy Number Variation

dbSNP : SNP database

FDR : taux de fausses découvertes (False Discovery Rate)

FWER : Family-Wise Error Rate

GSEA : Gene Set Enrichment Analysis

GWAS : étude d'association génome entier (Genome Wide Association Study)

IBD : Identical By Descent

kb : kilobase (1000 bp)

LD : Déséquilibre de Liaison (Linkage Desequilibrium)

MAF : Fréquence de l'Allèle Mineur (Minor Allele Frequency)

Mb : mégabase (10^6 bp)

NP : Non Progresseur

OR : Rapport de cotes (Odds Ratio)

PR : Progresseur Rapide

ROS : espèce réactive oxygénéé (Reactive Oxygen Species)

SIDA : Syndrome d'ImmunoDéficience Acquise

SNP : Single Nucleotide Polymorphism

SU.VI.MAX : SUpplémentation en VItamines et en Minéraux AntioXydants

UV : Ultra-Violet

VIH-1 : Virus de l'Immunodéficience Humaine de type I

Première partie

Introduction

1 Introduction à la génétique

Il existe une grande diversité parmi les caractères exprimés par les individus d'une même espèce. Chez les êtres humains, la couleur de peau, la taille ou le groupe sanguin sont autant d'exemples de cette diversité. Celle-ci se manifeste également dans la réponse de chaque individu face aux agents pathogènes. L'infection par le Virus de l'Immunodéficience Humaine de type I (VIH-1) est une parfaite illustration de ces différences. En effet, le Syndrome de l'ImmunoDéficience Acquise (SIDA), phase finale de l'infection par le VIH-1, apparaît en moyenne une dizaine d'années après avoir contracté le virus. Cependant, certaines personnes, appelées Progresseurs Rapides (PR), vont évoluer rapidement vers le SIDA en moins de 3 ans tandis que d'autres, les Non Progresseurs (NP), ne présentent aucun signe clinique du SIDA malgré une infection par le VIH-1 après plus de 10 ans.

La transmission des caractères se fait également d'une génération à l'autre pour les individus apparentés. Une fois de plus, la couleur de la peau, la couleur des yeux ou le groupe sanguin sont des exemples de caractères physiques héréditaires. Certaines maladies sont également héréditaires. Par exemple, si un parent est atteint d'une certaine pathologie, alors l'enfant sera lui aussi inévitablement atteint de la même affection, comme par exemple le syndrome de Leigh et le syndrome de Pearson. Dans d'autres cas, si un parent est atteint d'une maladie, l'enfant aura « seulement » une probabilité plus importante d'être atteint sans obligatoirement être malade, comme cela est le cas pour la mucoviscidose ou encore la drépanocytose.

Motivées par la compréhension de ces phénomènes et dans le but de mieux contrôler les conséquences des maladies, de nombreuses recherches ont permis de mettre en évidence une dimension génétique de ces variations. En 1953, Watson et Crick parachèvent la découverte d'une molécule à la base de l'hérédité et des variations du vivant : l'Acide Désoxyribo-Nucléique (ADN). Cette découverte va permettre l'avènement de nouvelles sciences telles que la génétique ou la génomique. La génétique étudie les parties de l'ADN, appelées gènes, responsables des caractères observables chez un individu. La génomique, quant à elle, ne se focalise pas uniquement sur un gène mais s'intéresse au fonctionnement d'un organisme à l'échelle de l'ensemble du matériel génétique, le génome.

1.1 L'Acide Désoxyribonucléique

Dans une cellule eucaryote, le génome d'un individu, soit l'ensemble de ses gènes, est réparti au sein du noyau en molécules compactes d'ADN appelées chromosomes. Un être humain possède 23 paires de chromosomes : 22 paires d'autosomes numérotées de 1 à 22 par taille décroissante et une paire de chromosomes sexuels (XX pour les femmes et XY pour les hommes). Chaque molécule d'ADN est constituée d'une succession de molécules appelées nucléotides. Un nucléotide est formé par l'association d'une base azotée, d'un désoxyribose et d'un groupement phosphate. Il existe quatre nucléotides différents : l'Adénine (A), la Cytosine (C), la Guanine (G) et la Thymine (T).

Durant la seconde moitié du XXème siècle, les nombreuses recherches sur l'ADN vont permettre la découverte de sa structure. La première propriété majeure a été énoncée en 1949 par le chimiste autrichien Erwin Chargaff [1, 2]. Elle stipule que la quantité de A et de C dans n'importe quel segment d'ADN est sensiblement égale à celle de T et de G respectivement, seule la quantité de A et de C (et donc de T et de G) diffère. Ce résultat suggère alors une complémentarité des nucléotides. La seconde propriété, publiée en 1953 par James Watson et Francis Crick [3], établit que l'ADN a une structure hélicoïdale à double brins (Figure 1). Cette structure et la complémentarité des nucléotides assurent une stabilité cruciale à l'ADN permettant une réplication fiable de cette molécule lors de la division cellulaire.

Les deux brins de l'ADN étant complémentaires, il suffit donc de ne connaître qu'un seul de ces deux brins pour disposer de la totalité de l'information génétique. Ainsi, cet acide est représenté par une séquence non aléatoire de nucléotides correspondant à un des deux brins.

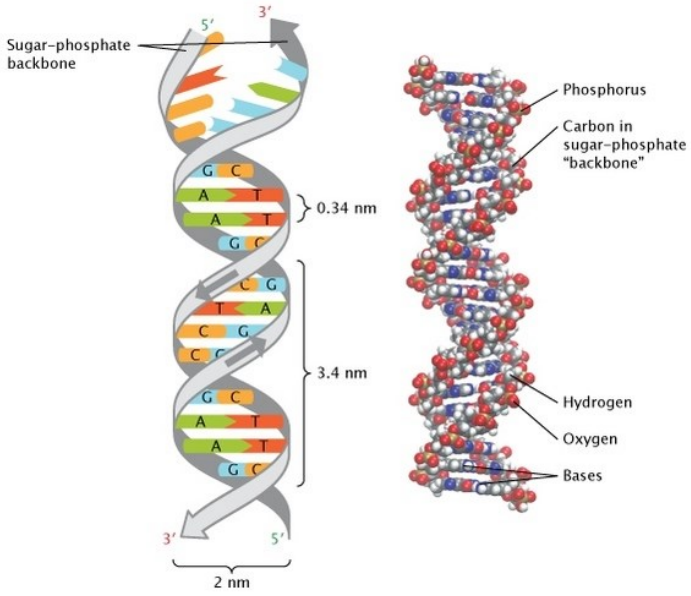


Figure 1. Représentation d'une portion d'ADN. Les nucléotides sont appariés selon leur complémentarité (A/C et G/T). Les deux brins complémentaires s'entrelacent pour former une double hélice. (Tiré de Pray, L. (2008) Discovery of DNA structure and function: Watson and Crick. Nature Education 1(1):100).

1.1.1 L'ADN est le support de l'information génétique

L'ADN contient toutes les informations nécessaires à la formation et à la vie de l'organisme. En effet, selon le dogme central de la biologie moléculaire, certaines portions de l'ADN, les gènes, sont à l'origine de la synthèse des protéines. Les régions codantes des gènes, dites exons, sont transcrites en Acide RiboNucléique messagers (ARNm) qui sont ensuite traduits en protéines (Figure 2). L'ARNm correspond au brin complémentaire de la séquence d'ADN transcrite à l'exception du T qui est remplacé par l'Uracile (U). Lors de la traduction, la succession de codons, triplets de nucléotides de l'ARNm, définit la succession d'acides aminés formant la protéine. En effet, à chaque codon correspond un acide aminé. Cette table de correspondance s'appelle le code génétique.

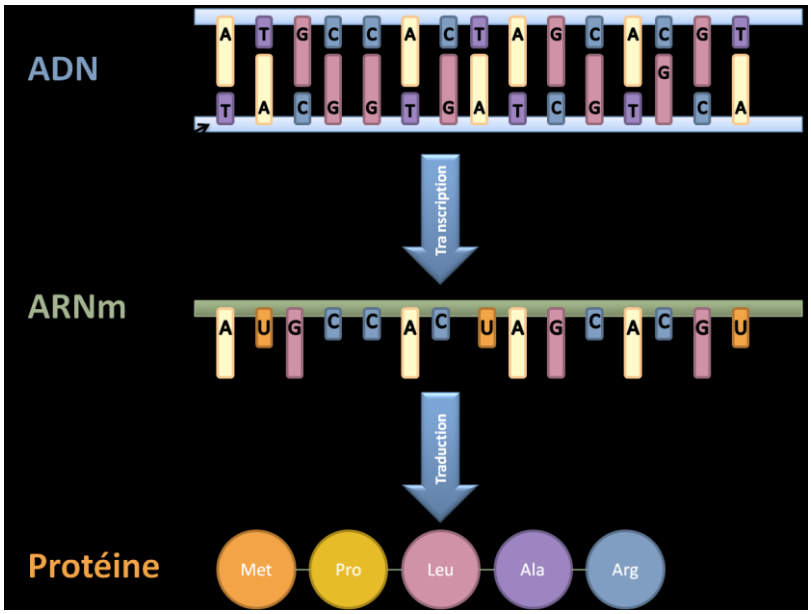


Figure 2. Représentation de la synthèse des protéines. La séquence d'ADN est transcrite en ARN messager. L'ARN messager est ensuite traduit en protéine.

Jusqu'à présent, environ 25 000 gènes ont été recensés chez l'Homme mais ils ne constituent que 1,5% des 3 milliards de nucléotides formant l'ADN. Les fonctions des longues régions non codantes sont encore peu connues mais il a été établi que ces régions, et notamment les parties proches de gènes, peuvent intervenir dans des mécanismes de régulation de la synthèse des protéines.

1.1.2 L'ADN est aussi le support de l'hérédité

L'ADN joue également un rôle majeur dans l'hérédité. Chaque individu est diploïde, c'est-à-dire qu'il possède une paire de chaque chromosome dans le noyau de ses cellules car il a hérité d'un chromosome de son père et d'un chromosome de sa mère. En effet, les cellules permettant la procréation, aussi appelées gamètes, présentes chez le père (spermatozoïde) comme chez la mère (ovule), ne possèdent qu'un seul exemplaire de chaque chromosome. Lors de la fécondation, c'est-à-dire la rencontre entre deux gamètes du sexe opposé, les chromosomes contenus dans chacune de ces deux cellules sont mis en commun et le fœtus possède donc une paire de chaque chromosome dont un est issu du père et l'autre de la mère. Ce mode de fonctionnement permet un brassage génétique puisque quatre combinaisons sont alors possibles (Figure 3).

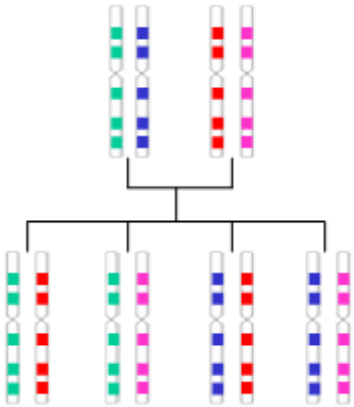


Figure 3. A partir des deux paires de chromosomes parentaux (en haut), il est possible d'observer 4 combinaisons différentes chez l'enfant (en bas).

La recombinaison génétique, aussi appelée « crossing-over », est un phénomène naturel qui se produit lors de la méiose. Au cours de cette division cellulaire à l'origine des gamètes, il se peut qu'au sein d'une paire chromosomique se produise un échange d'un segment d'ADN entre les deux chromosomes homologues (Figure 4).

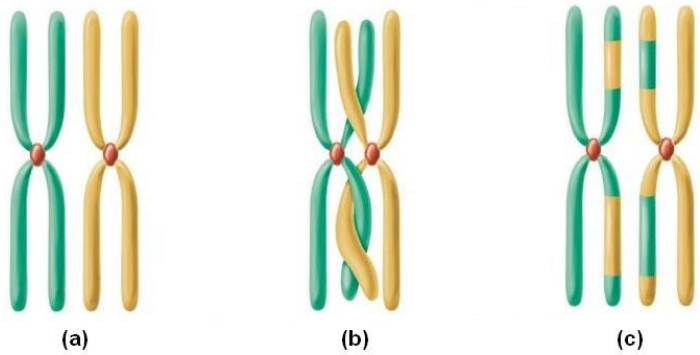


Figure 4. Schématisation du phénomène de recombinaison génétique pouvant intervenir durant la méiose. Au cours de cette division cellulaire, les chromosomes homologues s'assemblent en paires (a). Lors du rapprochement de ces chromosomes, des échanges de chromatides peuvent se produire (b) créant de nouveaux chromosomes (c).

La probabilité qu'une recombinaison à un endroit particulier d'un chromosome ait lieu est mesurée par le taux de recombinaison et varie en fonction des différentes paires de chromosomes mais aussi selon la position sur le chromosome. En effet, les extrémités des chromosomes ont plus de chance d'être soumises à ce phénomène que la partie centrale du chromosome. De plus, la probabilité qu'une recombinaison intervienne entre deux loci d'un même chromosome augmente avec la distance physique les séparant. Le « crossing-over »

permet donc l'apparition d'une nouvelle séquence génétique et contribue largement à la diversité génétique, et donc des caractères, observée chez les individus d'une même espèce.

Les séquences d'ADN présentent une diversité, à l'origine de celle des caractères observés, amplifiée par le brassage génétique dû à la reproduction et au phénomène de recombinaison. Les progrès techniques, et notamment le séquençage de l'ADN, ont permis la découverte de nombreux polymorphismes génétiques à l'origine de ces différences.

1.2 Les différents polymorphismes génétiques

Un polymorphisme génétique est une différence observée sur le même locus chromosomique chez des individus d'une même espèce. Chacune des versions de ce locus est appelée un allèle. Généralement, les polymorphismes apparaissent dans les cellules germinales en raison notamment d'erreurs dans la réplication de l'ADN et sont donc transmissibles d'une génération à l'autre. Ces modifications peuvent se produire à l'échelle d'un chromosome entier ou d'un nucléotide uniquement. Les différents allèles d'un gène peuvent soit n'avoir aucune conséquence sur la fonction de ce gène, soit en affecter la fonction selon trois modalités : perte de fonction, maintien partiel de la fonction avec interférences ou gain de fonction.

Chez les espèces diploïdes, c'est-à-dire possédant des paires de chromosomes, la combinaison des deux allèles observés à un locus particulier est appelée le génotype. Un individu diploïde dont les deux allèles d'un locus sont différents est dit hétérozygote. S'ils sont identiques, l'individu est dit homozygote pour l'allèle observé.

1.2.1 Les polymorphismes chromosomiques

Les polymorphismes chromosomiques sont des variations structurales résultant d'événements de translocation (échange réciproque de segment d'ADN entre des chromosomes non homologues) (Figure 5), d'inversion (renversement bout à bout d'un segment du chromosome), de fusion ou de fission de fragments chromosomiques. Des anomalies dans le nombre de chromosomes peuvent également être observées (Figure 5). Ces variations ne sont pas nécessairement liées à des anomalies phénotypiques.

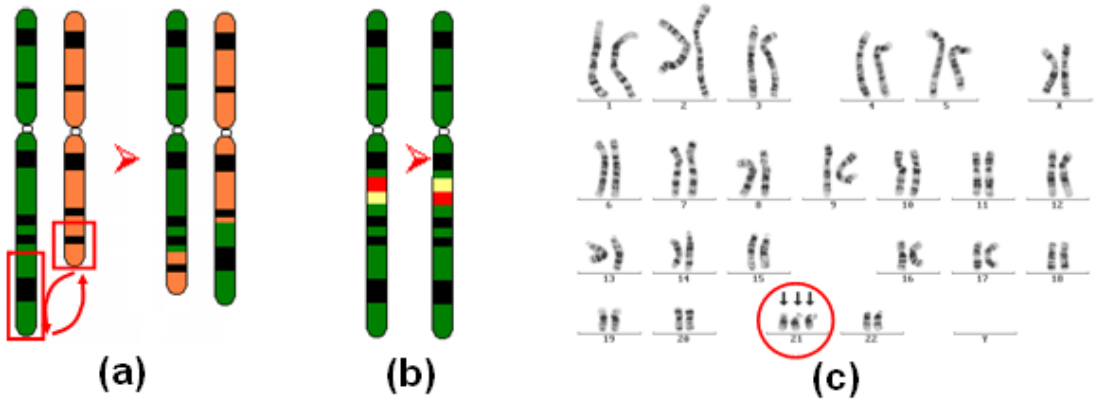


Figure 5. Exemples de polymorphismes chromosomiques. (a) Translocation entre 2 chromosomes. (b) Inversion au sein d'un même chromosome. (c) Anomalie du nombre de chromosomes : exemple de la trisomie 21.

1.2.2 Les séquences répétées en tandem

Les séquences répétées en tandem sont des répétitions consécutives d'un motif de nucléotides. Le nombre de répétitions peut, tout comme le nombre de nucléotides composant le motif répété, être de tailles variables (Figure 6). Plusieurs types de séquences répétées sont distingués selon ces paramètres. Les microsatellites sont des motifs de 1 à 5 nucléotides, ou paire de bases (bp), répétés de 2 à 50 fois consécutivement. Les mini-satellites sont des motifs plus grands, entre 15 et 100 bp répétés entre 15 et 50 fois. Enfin, les satellites sont constitués de motifs plus grands (α : 171, β : 168, et γ : 220 nucléotides respectivement) répétés un grand nombre de fois. Le nombre de répétitions du motif caractérise l'allèle de ce polymorphisme. Ces marqueurs ont longtemps été des marqueurs génétiques de prédilection.

...ATGCTAGATG $(CA)_n$ TGCCATAACG...

Figure 6. Exemple de séquence répétée en tandem. Le motif répété n fois est constitué de deux nucléotides. Il s'agit donc d'un microsatellite.

1.2.3 Les indels

Les indels correspondent à une insertion ou une délétion d'un ou plusieurs nucléotides dans la séquence d'ADN (Figure 7). La taille de l'insertion ou de la délétion est variable.

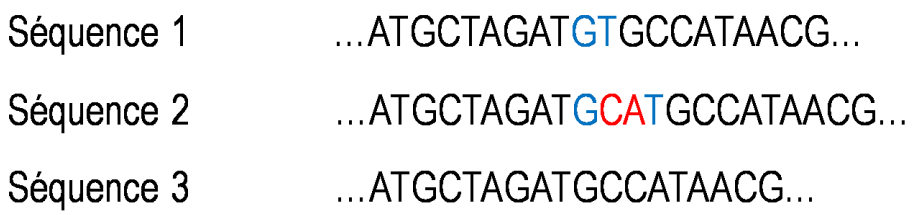


Figure 7. Exemples d'indels. La séquence 1 fait office de référence. La séquence 2 comporte une insertion de deux nucléotides (en rouge) entre le G et le T (en bleu) par rapport à la séquence 1. La séquence 3 illustre une délétion des deux nucléotides G et T (en bleu sur la séquence 1) par rapport à la séquence de référence.

1.2.4 Les Single Nucleotide Polymorphisms

Un Single Nucleotide Polymorphism (SNP) est le changement d'un seul nucléotide à un locus particulier (Figure 8). Ce polymorphisme est le type de variation le plus commun du génome et représente près de 90% des différences génétiques observées entre individus. D'après les dernières données du projet 1000 Génomes [4], environ 40 millions de SNPs ont d'ores et déjà été identifiés sur l'ensemble du génome humain. Etant donnée la facilité avec laquelle ils peuvent être caractérisés, les SNPs sont devenus des marqueurs génétiques privilégiés. Ainsi, des cartographies très denses de ces polymorphismes, comme la base de données dbSNP [5], ont été développées ces dernières années.

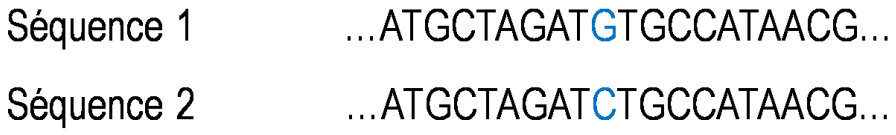


Figure 8. Exemple de SNP. Le nucléotide G (en bleu) de la séquence 1 est remplacé par un C (en bleu) dans la séquence 2.

1.2.5 Les Copy Number Variations

Les Copy Number Variations (CNVs) sont des polymorphismes découverts récemment et représentent près de 12% du génome. Ils sont définis comme étant des séquences d'ADN d'une longueur supérieure à 1kilobase (kb), soit 1000 nucléotides, répétées un nombre de fois variable d'un individu à l'autre. Les CNVs sont le résultat d'événements d'insertion, délétion et duplication. Ce nouveau type de polymorphisme suscite un grand intérêt au sein de la communauté scientifique et ouvre de nouvelles perspectives en génétique épidémiologique puisqu'il peut modifier le niveau d'expression d'un gène et être à l'origine de pathologies [6].

2 Notions de génétique des populations

2.1 Modèle de Hardy-Weinberg

En 1908, un mathématicien anglais, Godfrey Harold Hardy, et un médecin allemand, Wilhelm Weinberg, établissent indépendamment un résultat majeur de la génétique des populations. En effet, ils démontrent que, sous certaines conditions, les fréquences des allèles et des génotypes d'un polymorphisme bi-allélique au sein d'une population restent constantes au cours du temps [7, 8]. Ce résultat est connu comme étant l'équilibre d'Hardy-Weinberg.

2.1.1 Enoncé de l'équilibre d'Hardy-Weinberg

Considérons une population diploïde ***P*** dont la reproduction est sexuée et un polymorphisme ***S*** bi-allélique dont les allèles sont notés A et a. Notons respectivement f_A et f_a les fréquences de l'allèle A et de l'allèle a à la génération *t*.

Si les hypothèses suivantes sont vérifiées :

- La population ***P*** est d'effectif infini ou très grand.
- Les générations sont non chevauchantes.
- La panmixie est respectée dans ***P*** : les individus de la génération *t+1* sont obtenus par des croisements aléatoires entre les individus de la génération *t*, donc par tirages aléatoires de deux chromosomes dans ceux de ***P*** à la génération *t*.
- Il n'y a pas de nouvelle mutation au locus étudié, de phénomène de sélection naturelle et de migration.

alors, la fréquence des génotypes à la génération *t+1* est donnée par :

$$\begin{cases} f_{AA} = f_A^2 \\ f_{Aa} = 2f_A f_a \\ f_{aa} = f_a^2 \end{cases}$$

où f_{AA}, f_{Aa} et f_{aa} désignent respectivement les fréquences des génotypes AA, Aa et aa.

Etant donné que la somme des fréquences alléliques à la génération *t* est égale à 1, on vérifie facilement que la somme des fréquences génotypiques à la génération *t+1* est égale à 1. Enfin, si g_A et g_a désignent respectivement les fréquences de A et a à la génération *t+1* :

$$\begin{cases} g_A = f_{AA} + \frac{1}{2}f_{Aa} = f_A^2 + f_A f_a = f_A(f_A + f_a) = f_A \\ g_a = f_{aa} + \frac{1}{2}f_{Aa} = f_a^2 + f_A f_a = f_a(f_A + f_a) = f_a \end{cases}$$

Par récurrence, on en déduit que les fréquences des allèles A et a restent inchangées au cours des générations.

2.1.2 Influence des hypothèses

En raison des nombreuses hypothèses sous-jacentes, le modèle d'Hardy-Weinberg devrait apparaître comme théorique, abstrait et irréaliste. Cependant, l'étude de la plupart des gènes donne des résultats compatibles avec ce modèle. Ce paradoxe s'explique par le fait que certaines des conditions précédentes peuvent être parfaitement réalisées et d'autres, bien qu'illégitimes, n'ont d'effet perceptible que sur une longue échelle de temps et peuvent donc être négligées sur l'espace de quelques générations.

La panmixie est la condition la plus facilement réalisée. Dans la plupart des espèces, les unions et les fécondations sont aléatoires pour la plupart des gènes. Par exemple, chez l'Homme, les unions ne sont pas panmictiques pour les gènes gouvernant la taille ou la pigmentation mais elles le sont pour les gènes gouvernant les groupes sanguins, les facteurs sériques ou les maladies. Les conséquences des mutations ou des pressions de sélection ne sont quant à elles visibles que sur des centaines ou des milliers de générations.

Les hypothèses les plus délicates concernent l'absence de migration et la taille infinie de la population. L'effet des migrations peut modifier la constitution génétique d'une population en quelques générations.

Par essence, la taille infinie d'une population est une condition impossible. Les populations finies sont le siège d'un phénomène appelée la dérive génétique, c'est-à-dire la fluctuation aléatoire des fréquences des allèles. Une population de petite taille sera plus soumise aux biais d'échantillonnage lors des unions, ce qui conduira à une grande fluctuation des fréquences alléliques (Figure 9). Une population sera considérée comme infinie lorsque les effets de la dérive génétique sont négligeables.

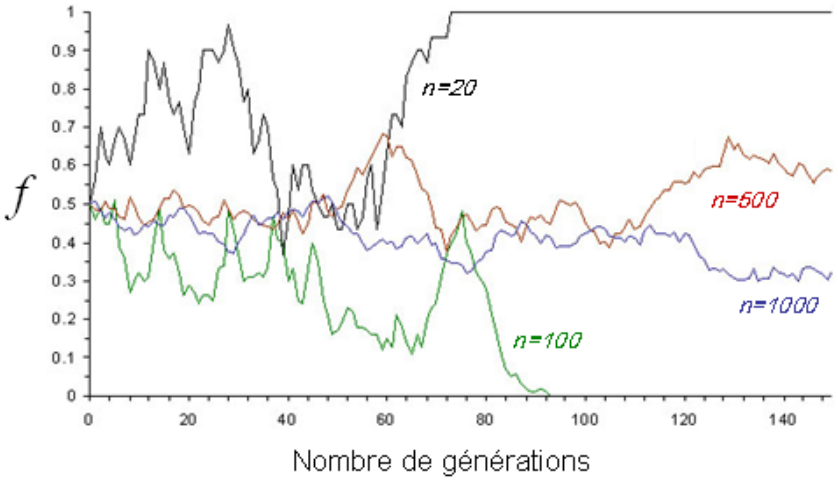


Figure 9. Modélisation de la dérive génétique pour différentes tailles de population. La fréquence allélique d'origine est 0,5. Les amplitudes des fluctuations de la fréquence allélique f sont inversement proportionnelles aux tailles des populations. La population la plus grande (en bleu) se rapproche le plus de l'équilibre d'Hardy-Weinberg.

Lorsqu'un polymorphisme dévie de l'équilibre de Hardy-Weinberg, il convient d'en identifier les raisons. Elles peuvent être, entre autres, démographiques (migration de population, réduction importante et rapide de l'effectif d'une population, écart à la panmixie) ou liées à des événements de sélection naturelle.

2.2 Déséquilibre gamétique et déséquilibre de liaison

L'étude de la composition génétique d'une population et de son évolution pour un seul gène est très restrictive, d'autant plus que de nombreux caractères ou de nombreuses pathologies ne dépendent pas uniquement d'un gène mais de plusieurs.

L'étude de l'évolution simultanée de plusieurs gènes devient rapidement si complexe qu'elle ne permet pas de mise en équation comme pour l'équilibre de Hardy-Weinberg, mais s'avère d'un grand intérêt. En effet, l'analyse de cette situation permet d'introduire le concept de déséquilibre gamétique, d'une grande utilité pour élaborer une cartographie des gènes, analyser l'origine de certaines mutations mais également pour élaborer de nouveaux marqueurs diagnostics de risque génétique.

2.2.1 Définition du déséquilibre gamétique

Lorsque l'étude de la diversité génétique se focalise sur un gène, la fréquence d'un allèle est égale à la fréquence du gamète portant cet allèle. Cette égalité ne tient plus quand deux gènes ou plus sont étudiés simultanément.

Considérons :

- un gène A dont les allèles sont A_1 de fréquence p et A_2 de fréquence q
- un gène B dont les allèles sont B_1 de fréquence u et B_2 de fréquence v

Quatre types de gamètes différents portant chacun une combinaison d'allèles de chaque gène sont potentiellement observables :

- le gamète (A_1, B_1) de fréquence f_{11}
- le gamète (A_1, B_2) de fréquence f_{12}
- le gamète (A_2, B_1) de fréquence f_{21}
- le gamète (A_2, B_2) de fréquence f_{22}

Nécessairement, il existe une relation entre les fréquences gamétiques et les fréquences alléliques de chaque gène, bien que celle-ci ne soit pas évidente.

Supposons que les allèles des deux gènes sont « réunis » indépendamment les uns les autres et aléatoirement dans les gamètes. Cette hypothèse conduit alors à une situation appelée équilibre gamétique et les fréquences de chaque gamète sont égales au produit des fréquences des allèles qu'il porte :

$$\begin{cases} f_{11} = pu \\ f_{12} = pv \\ f_{21} = qu \\ f_{22} = qv \end{cases}$$

Néanmoins, même si la population est à l'équilibre de Hardy Weinberg pour chacun de ces deux gènes, cette situation d'équilibre gamétique ne revêt ni un caractère obligatoire ni un caractère courant. Le non respect de ces égalités est appelé le déséquilibre gamétique, défini

comme la différence entre la fréquence réelle d'un gamète et sa fréquence théorique à l'équilibre :

$$\begin{array}{lll} pour & (A_1, B_1) & \Delta_{11} = f_{11} - pu \\ & (A_1, B_2) & \Delta_{12} = f_{12} - pv \\ & (A_2, B_1) & \Delta_{21} = f_{21} - qu \\ & (A_2, B_2) & \Delta_{22} = f_{22} - qv \end{array}$$

Tous les mécanismes supposés inexistants dans le modèle d'Hardy Weinberg, tels que les migrations de population ou les mutations, peuvent être à l'origine d'un déséquilibre gamétique.

2.2.2 Evolution temporelle du déséquilibre gamétique

D'un point de vue génétique, le déséquilibre gamétique est fortement dessiné par les phénomènes de recombinaisons. Ainsi, deux gènes proches sur un même chromosome sont plus enclins à être dans une telle configuration. Pour cette même raison et à force de recombinaisons, ce déséquilibre tend à disparaître avec le temps pour retourner vers l'équilibre gamétique.

En utilisant les mêmes notations qu'au paragraphe précédent, il et possible de quantifier le temps nécessaire à la disparition du déséquilibre gamétique entre les gènes A et B en fonction du taux de recombinaison *r* entre ces deux gènes.

Notons $f_{11,\ i-1}$ la fréquence du gamète (A_1, B_1) à la génération *i-1*. A la génération *i*, ces gamètes seront issus de deux phénomènes :

- les gamètes (A_1, B_1) de la génération *i-1* qui n'ont pas recombiné. La probabilité d'occurrence de cet événement est *(1 – r)*.
- les nouveaux gamètes (A_1, B_1) issus de recombinaisons associant l'allèle A_1 (de fréquence p) et l'allèle B_1 (de fréquence u).

La fréquence de ce gamète à la génération *i* s'écrit alors :

$$f_{11,i} = (1 - r) f_{11,i-1} + rpu$$

Cette équation, après avoir retranché *pu* à ses deux membres, équivaut à :

$$\Delta_{11,i} = (1 - r)\Delta_{11,i-1}$$

d'où,

$$\Delta_{11,i} = (1 - r)^i \Delta_0$$

Cette dernière équation prouve que le déséquilibre gamétique tend vers 0 avec le temps. La vitesse de cette décroissance dépend seulement de *r* et donc de la liaison génétique ou non existante entre les gènes A et B. Ainsi, si les gènes ne sont pas liés, ce déséquilibre gamétique va disparaître très rapidement. En revanche, si les gènes sont très liés génétiquement (*r* faible), le déséquilibre pourra perdurer au cours du temps (Figure 10). Le complexe majeur d'histocompatibilité est un exemple de déséquilibre persistant au fil des générations.

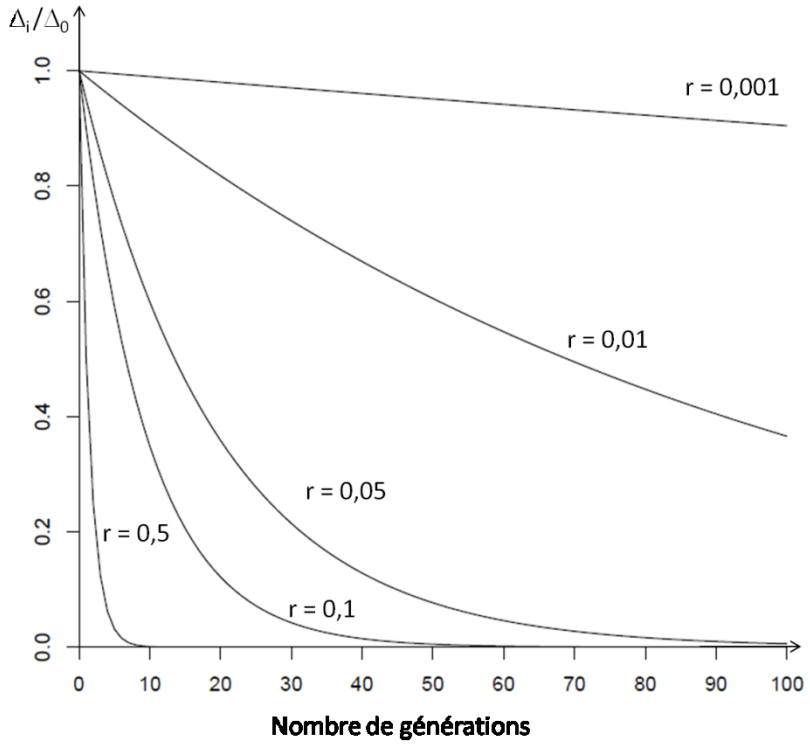


Figure 10. Vitesse de disparition du déséquilibre gamétique pour différentes valeurs du taux de recombinaison entre deux loci. Plus le taux de recombinaison est grand, plus le déséquilibre disparaît rapidement.

2.2.3 Déséquilibre de liaison

Un déséquilibre gamétique entre deux loci peut perdurer en raison d'une importante liaison génétique entre ceux-ci. Les Anglo-Saxons ont alors qualifié cette situation de *linkage*

disequilibrium, traduit en français par déséquilibre de liaison (LD). Ainsi, le déséquilibre de liaison entre deux loci suggère qu'il existe une liaison génétique et un déséquilibre génétique.

Cependant, cette formule est souvent employée à tort. En effet, une liaison génétique entre deux gènes peut exister sans qu'un déséquilibre gamétique ne puisse être observé, et inversement. L'observation d'un déséquilibre gamétique n'est pas nécessairement l'indication d'une liaison génétique. Cette confusion est fréquente dans de nombreuses études d'épidémiologie génétique.

Dans la suite de cette thèse, par souci de simplicité, nous assumons volontairement la confusion entre ces deux notions. De plus, nous travaillons sur de petites régions chromosomiques dans lesquelles la liaison génétique semble pertinente.

2.2.4 Les différentes mesures du déséquilibre de liaison

Plusieurs mesures du déséquilibre de liaison ont été développées à partir de la comparaison des fréquences alléliques et des fréquences des couples d'allèles. En utilisant les mêmes notations que précedemment, la première mesure introduite est notée *D* [9] :

$$\begin{cases} f_{11} &= pu + D \\ f_{12} &= pv - D \\ f_{21} &= qu - D \\ f_{22} &= qv + D \end{cases}$$

L'équilibre gamétique se traduit par un coefficient *D* égal à 0. Cette mesure n'est que très rarement utilisée en pratique. En effet, elle est dépendante des fréquences alléliques et la comparaison de cette mesure pour deux couples de polymorphismes n'est pas aisée. Afin de pallier cet inconvénient, une mesure normalisée, notée *D'*, a été proposée [10] :

$$D' = \begin{cases} \frac{D}{min(pu,qv)} & pour \quad D \geq 0 \\ \\ \frac{D}{min(pv,qu)} & pour \quad D \leq 0 \end{cases}$$

Ce coefficient *D'* varie entre -1 et 1. Lorsqu'il vaut 0, l'équilibre de liaison est vérifié. Quand il est égal à 1 ou -1, cela signifie qu'une ou deux combinaisons d'allèles ne sont pas observées au sein de la population. Cependant, cette mesure est encore une fois dépendante des

fréquences. La mesure la plus utilisée permettant de s'affranchir de cette dépendance aux fréquences est le r^2 [11] :

$$r^2 = \frac{D^2}{pquv}$$

Ce coefficient varie de 0 à 1 : un r^2 égal à 0 indique une situation d'équilibre alors qu'un r^2 égal à 1 indique une situation de déséquilibre de liaison total. Le déséquilibre de liaison total désigne la situation où les allèles des deux SNPs sont parfaitement corrélés et systématiquement co-transmis. Dans ce cas, la connaissance du génotype d'un SNP détermine totalement le génotype de l'autre SNP.

2.3 Haplotypes

La notion de déséquilibre de liaison est adaptée pour l'étude simultanée de deux loci mais n'est pas la plus pertinente lorsqu'une région chromosomique contenant plus de deux polymorphismes est considérée. Il est naturellement possible d'évaluer le LD en considérant les paires de SNPs de la région mais cette méthode ne permet pas de capturer pleinement la structure complexe des corrélations entre les allèles de tous les SNPs de la région. Dans une telle situation, l'étude des haplotypes est privilégiée.

Un haplotype est défini comme la combinaison d'allèles de deux SNPs ou plus sur le même chromosome (Figure 11).

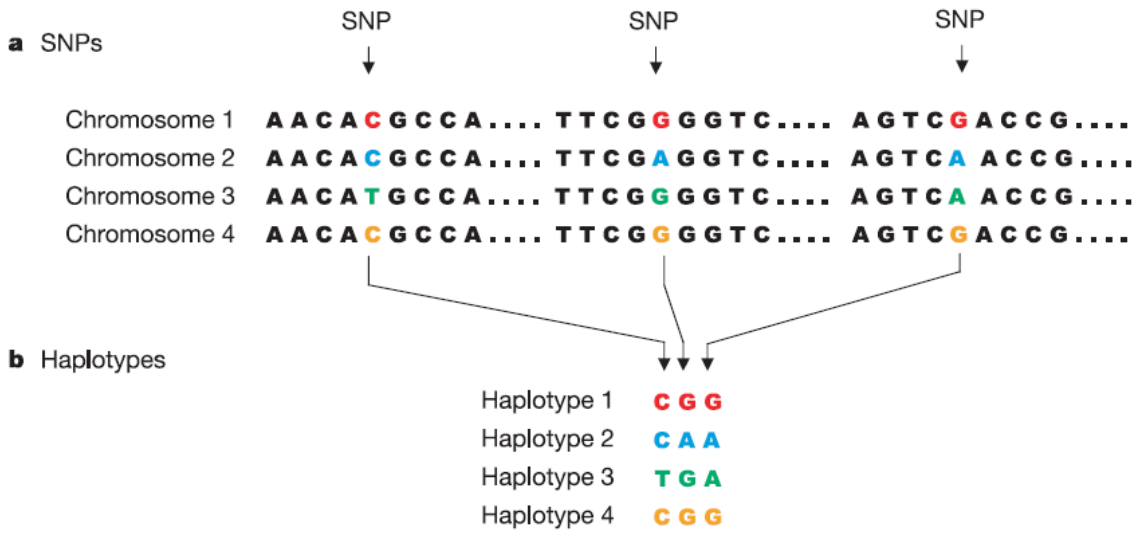


Figure 11. Exemple d'haplotypes composés de 3 SNPs. (Adaptée de [12])

Généralement, les haplotypes sont créés par une succession de mutations dont les combinaisons d'allèles sont brassées ou non par différentes recombinaisons (Figure 12). La disparition ou non de ces haplotypes dépend de divers paramètres tels que la dérive génétique, la sélection naturelle ou les migrations.

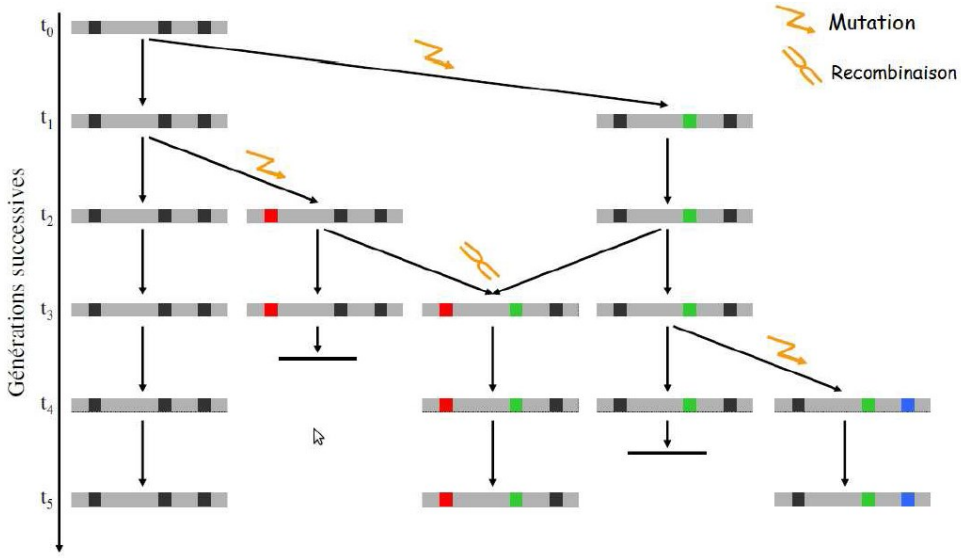


Figure 12. Exemple de génèse d'haplotypes dans une région de 3 SNPs.

Le nombre de combinaisons observables croît avec le nombre de SNPs considérés et dépend également des taux de recombinaison de la région chromosomique. En effet, en s'intéressant à *s* SNPs, il est possible d'observer jusqu'à 2^s haplotypes et de forts taux de recombinaison dans la région étudiée favorisent l'apparition de nouvelles combinaisons. Cependant, des études ont montré que les taux de recombinaison ne sont pas uniformes le long du génome et que les forts taux de recombinaison se concentrent généralement dans de petites régions séparées par des zones plus ou moins grandes dans lesquelles ces taux sont très faibles [13, 14]. Ces variations dans la distribution des taux de recombinaison expliquent la présence de « blocs » d'haplotypes. Au sein de ces blocs, le LD entre les SNPs est fort et la diversité haplotypique est relativement limitée du fait de l'absence de recombinaisons [15]. Au contraire, les déséquilibres de liaison entre deux blocs sont faibles et la diversité haplotypique est plus importante à cheval sur plusieurs blocs puisque les recombinaisons brassent les combinaisons d'allèles.

La notion d'haplotype est utilisée pour l'étude de nombreux phénomènes et notamment l'évolution démographique de l'espèce humaine. En effet, leur analyse permet de retracer l'histoire migratoire des populations ou de repérer des événements de sélection naturelle. Les

haplotypes ont également été utilisés en génétique épidémiologique afin de détecter des combinaisons d'allèles impliquées dans la susceptibilité à certaines pathologies [16, 17].

3 Introduction à la génétique épidémiologique

Au-delà de la recherche d'une meilleure connaissance du fonctionnement du corps humain, de l'évolution et de l'adaptation à l'environnement, la génétique permet d'appréhender la compréhension des maladies sous un angle nouveau lié à leur diffusion au sein de la population et des familles. Ce volet de la génétique est appelé génétique épidémiologique et a connu un essor considérable en raison notamment de l'amélioration des techniques de biologie moléculaire pour analyser l'ADN et les variations génétiques. La génétique épidémiologique joue un rôle important dans le processus de mise en cause d'un gène dans une maladie avec pour objectif final de développer et mettre en œuvre des outils préventifs, diagnostiques, et thérapeutiques [18].

3.1 Polymorphismes génétiques et pathologies

L'influence de la génétique sur le développement ou la réponse à certaines pathologies est désormais un phénomène connu. Les polymorphismes génétiques sont associés à des modifications de la structure des protéines ou de leur niveau d'expression et par conséquent à une perte de fonction plus ou moins importante. De nombreux polymorphismes génétiques associés à des pathologies ont d'ores et déjà été identifiés. Par exemple, la mucoviscidose peut être due à une délétion dans un exon du gène *CFTR* [19] et la maladie de Huntington a été associée à des répétitions en grand nombre d'un microsatellite localisé dans le premier exon du gène *HTT* [20]. Cependant, de nombreuses maladies communes comme les cancers, les diabètes ou les maladies auto-immunes entre autres ne sont pas seulement causées par une seule mutation génétique. Pour ces maladies dites multifactorielles, plusieurs facteurs génétiques peuvent exister et ils sont difficiles à identifier puisqu'ils sont nombreux et interagissent de plus avec des facteurs environnementaux. Par exemple, le gène *ApoE* a été identifié dans la maladie d'Alzheimer [21], les gènes *BRCA1* et *BRCA2* dans le cancer du sein [22] ou encore les gènes de la région *HLA* dans le SIDA [23].

3.2 Les différents types d'études génétiques et génomiques

Différents types d'études ont permis la découverte de facteurs génétiques impliqués dans des pathologies. La conception d'une telle étude doit prendre en compte deux paramètres : la

population considérée et les régions chromosomiques étudiées. De même, la nature du caractère étudié, appelé phénotype, conduit à des analyses différentes.

Bien que les résultats obtenus avec ces analyses aient permis la découverte de nombreuses liaisons ou associations avec des maladies, il est nécessaire de garder à l'esprit que les polymorphismes mis en évidence lors des analyses de génétique épidémiologique ne sont pas nécessairement les variants causaux de la maladie. Ceci peut s'expliquer par exemple par la non inclusion du variant causal dans l'étude ou parce que le polymorphisme causal est d'une nature différente des marqueurs étudiés. Ces analyses peuvent mettre en exergue une association entre la maladie et un variant en déséquilibre de liaison avec le variant causal (Figure 13). Ainsi, les conclusions de telles études doivent être prudentes et prendre en compte cette possibilité.

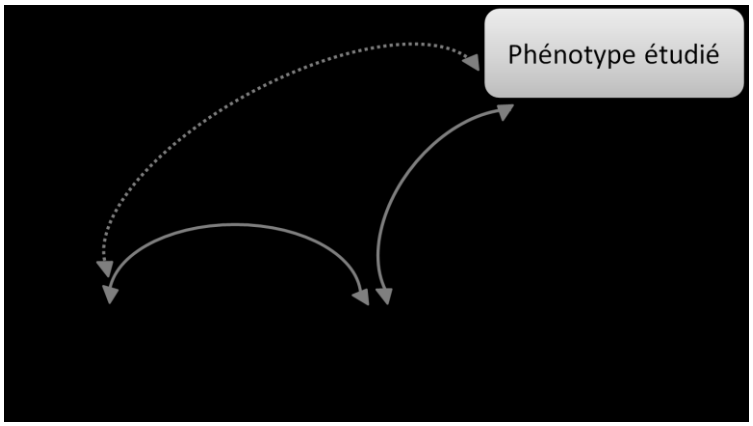


Figure 13. Découverte d'un variant associé à un phénotype. Le variant détecté lors d'une étude de génétique épidémiologique n'est pas nécessairement le variant causal de la maladie mais peut être en déséquilibre avec celui-ci. (Adaptée de [24])

3.2.1 Etudes de liaison et études d'association

Ces deux types d'études se distinguent par la population analysée. Les premières étudient l'histoire familiale d'une maladie alors que les secondes étudient la maladie au sein d'une population généralement non apparentée.

3.2.1.1 Etudes de liaison

Les études de liaison sont les premières à avoir vu le jour. Elles étudient la co-ségrégation des allèles d'un ou plusieurs polymorphismes et de la maladie au cours des générations au sein de familles (Figure 14).

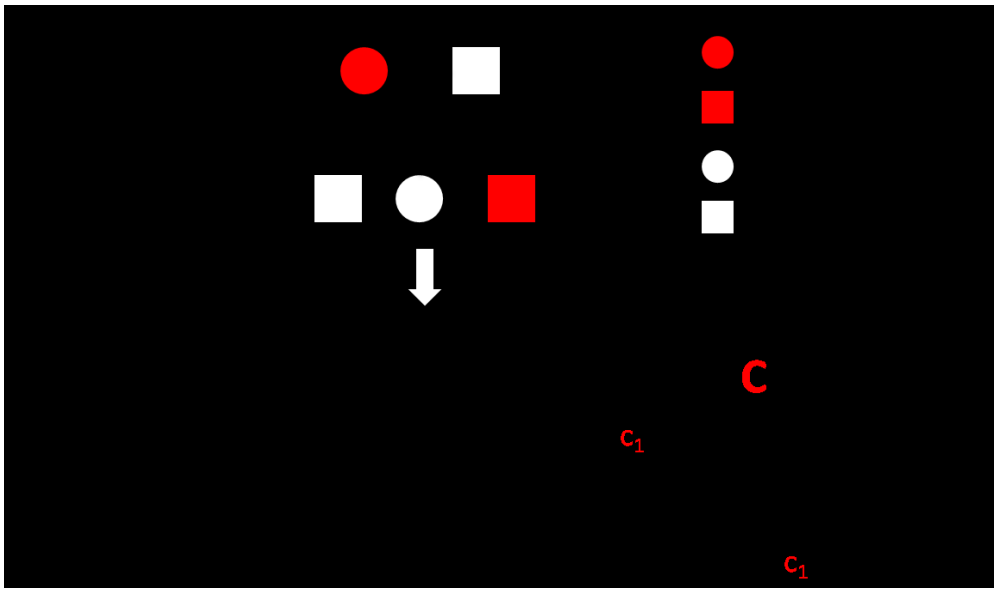


Figure 14. Exemple d'une généalogie incluant des individus malades (représentés en rouge) et des individus sains (représentés en blanc). La transmission de la maladie est étudiée pour 3 SNPs (A, B et C). Pour les SNPs A et B, aucun allèle n'est corrélé avec la présence ou l'absence de la maladie. A l'opposé, pour le SNP C, l'allèle c_1 est présent chez tous les individus malades, et absent chez tous les individus sains, suggérant son implication dans la survenue de la maladie.

Ce type d'étude familiale a notamment permis la découverte de facteurs génétiques responsables de maladies monogéniques telles que la mucoviscidose [25, 26] ou la maladie de Huntington [27-29]. Ces études ont abouti à des résultats intéressants concernant des maladies infectieuses telles que la lèpre [30]. Elles restent cependant limitées pour la détection des facteurs génétiques impliqués dans les maladies multifactorielles, comme le SIDA par exemple, pour lesquelles chaque facteur n'explique qu'une fraction du phénotype et les corrélations familiales sont plus difficiles à caractériser. De plus, ces pathologies ne touchent pas nécessairement plusieurs membres d'une même famille et la collecte d'information familiale perd donc de son intérêt. Pour ce type de maladies, les généticiens ont recours aux études d'association.

3.2.1.2 Etudes d'association

Ce type d'étude vise à détecter des variants génétiques associés à la maladie au sein d'une population d'individus non apparentés et non plus familiale comme les études de liaison. Bien que tous les individus soient apparentés à un certain degré, les études d'associations considèrent d'une part que les relations de parenté entre les individus sont inconnues, et

d'autre part, que ces liens sont lointains et remontent à un nombre suffisamment important de générations. Ainsi, des frères, parents, grands-parents ou des cousins ne sont pas inclus dans ce type d'étude.

Les méthodes visant à repérer de tels marqueurs dépendent de la nature du phénotype étudié. Classiquement, ces études comparent les distributions alléliques ou génotypiques entre une population d'individus affectés, appelés « cas », et des individus sains appelés « témoins ». Ce type d'étude, dit « cas/témoins », se focalise sur des phénotypes dichotomiques. Cependant, si le phénotype étudié est quantitatif, comme la tension artérielle ou la glycémie par exemple, d'autres approches statistiques sont employées mais le principe reste inchangé. Les techniques employées selon le phénotype étudié seront décrites plus précisément ultérieurement.

Le recrutement de patients à inclure dans les études d'associations s'avère a priori plus simple du fait que les individus ne sont pas apparentés. Néanmoins, de nombreuses précautions doivent être prises lors du recrutement de ces patients afin d'éviter les facteurs de confusion qui associeraient un polymorphisme non pas à cause d'une réalité génétique mais du fait de biais dans la population étudiée. En effet, des différences d'âge, de sexe ou d'origine ethnique peuvent conduire à la détection d'association sans fondement biologique (Figure 15).

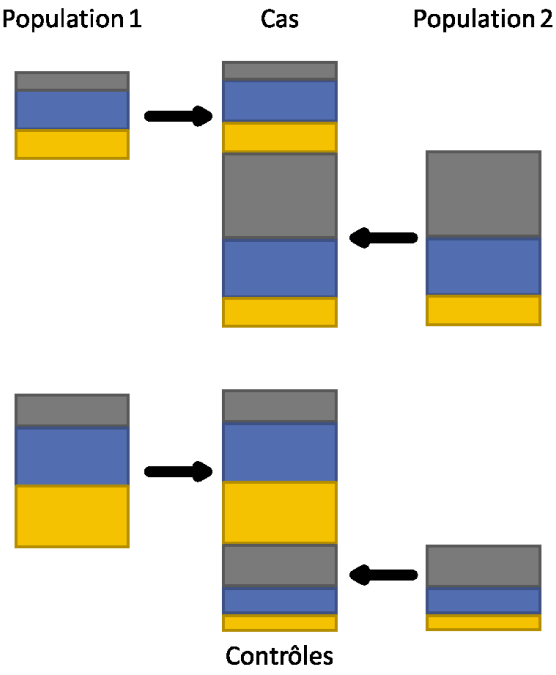


Figure 15. Exemple de structure de population dans une étude d'association de type « cas/contrôles ». (Adaptée de [24])

Ce genre d'étude a permis l'identification de gènes impliqués dans la pathogénèse ou la prédisposition à de nombreuses maladies multifactorielles. Par exemple, des études d'association ont mis en évidence le rôle de nombreux gènes dans des maladies multifactorielles comme la maladie de Parkinson, les cancers ou le SIDA [23, 31-35].

3.2.2 Etude gène ou SNPs candidats et étude génome entier

Le nombre de polymorphismes analysés dans les études de liaison ou d'association a considérablement augmenté durant les dernières années grâce à l'amélioration des techniques de biologie moléculaire. Originellement, les études se focalisaient uniquement sur quelques polymorphismes ou gènes dits « candidats », choisis pour leur lien connu ou supposé avec une maladie mais la dernière décennie a permis l'émergence des études « génome-entier » dans lesquelles l'ensemble des SNPs du génome est considéré.

3.2.2.1 Etudes gène candidat

Les approches gène-candidat sont les premières à avoir été réalisées en épidémiologie génétique et se focalisent sur l'étude de gènes connus ou potentiellement impliqués dans l'étiologie d'une maladie [36]. Le choix des gènes est généralement guidé par des a priori biologiques tels que la fonction du gène ou l'appartenance à une voie de signalisation particulière. De même, les résultats d'études publiées peuvent influencer le choix des marqueurs retenus dans ce type d'analyse. Généralement, ces études se focalisent sur un faible nombre de polymorphismes génétiques mais rien n'empêche qu'un grand nombre de gènes soient testés. De plus, étant donné le faible nombre de marqueurs inclus dans l'étude, il est possible d'effectuer un séquençage exhaustif des gènes considérés et d'avoir une information plus précise de la région étudiée qu'en utilisant des techniques de génotypage classiques. Par conséquent, l'influence de plusieurs types de polymorphismes comme les microsatellites ou les indels peut être analysée et de nouveaux polymorphismes peuvent être découverts. Les approches « gène-candidat » sont bien adaptées pour des maladies monogéniques mais sont moins puissantes pour découvrir l'ensemble des variants génétiques intervenant dans des maladies multifactorielles. En effet, en plus de l'absence de connaissance de facteurs potentiels et donc d'a priori, tous les polymorphismes génétiques impliqués dans l'étiologie de la maladie ou dans la réponse de l'individu à la pathologie peuvent ne pas être inclus dans cette étude. Néanmoins, ces stratégies ont permis l'identification de gènes impliqués dans différentes maladies telles que le SIDA [37, 38] ou la leucémie [39] par exemple.

3.2.2.2 Etudes génome entier

Au contraire des études gène-candidat, les approches génome entier ne se concentrent pas uniquement sur certains gènes mais sur l'ensemble des polymorphismes communs du génome. Ce type d'étude, s'affranchissant de tout a priori biologique, a donc pour but de repérer de nouveaux facteurs impliqués dans des maladies multifactorielles. Cependant, elles ne permettent pas obligatoirement la compréhension complète du rôle joué par les marqueurs repérés dans la pathogénèse de la maladie. Des études plus spécifiques (et notamment de type gène candidat) sont ensuite nécessaires pour améliorer cette compréhension.

L'amélioration des techniques de génotypage a largement contribué au développement de cette stratégie. Ainsi, plusieurs milliers d'études génome entier ont été réalisées au cours de la dernière décennie, comme le confirme la base de données GWAS Catalog recensant les résultats de ces types d'analyses [40]. Les études d'association génome entier (Genome Wide Association Study ou GWAS en anglais) seront détaillées plus précisément ultérieurement.

3.2.3 Etude longitudinale et étude transversale

La nature du phénotype analysé influe sur les méthodes utilisées et conduit à des analyses différentes.

Le phénotype étudié peut être une variable mesurée à un instant précis reflétant alors une situation figée. Les études s'intéressant à ce genre de phénotypes sont appelées études transversales. Ces analyses établissent un « cliché » de la population à un moment donné et visent à déterminer des associations entre un polymorphisme génétique et l'exposition à une maladie. Les méthodes statistiques employées pour rechercher de telles associations sont nombreuses : tests d'indépendance, régression linéaire…

Le phénotype peut également être une variable mesurant une durée avant la survenue d'un événement comme la mort ou la rémission d'une tumeur par exemple. Ces analyses de suivi longitudinal sont plus délicates à mettre en œuvre car elles nécessitent un suivi régulier des patients mais permettent de rechercher des marqueurs génétiques impliqués dans des maladies longues comme le SIDA [41-43], les cancers [44-46] ou les maladies cardio-vasculaires [47, 48]. Lors de ces études, des modèles de survie, telle que la régression de Cox, sont utilisés [49].

4 Outils génomiques et bioinformatiques

L'essor des études d'association génome-entier ou GWAS (Genome Wide Association Study en anglais) a été rendu possible grâce au progrès concommitants des technologies de biologie moléculaire et de l'informatique.

4.1 Le génotypage et les puces à ADN

Le génotypage vise à déterminer les génotypes d'un polymorphisme chez un individu. Des puces à ADN, permettant de génotyper un nombre croissant de SNPs ont été développées depuis 10 ans. Schématiquement, une puce est une plaque sur laquelle se trouvent des microbilles, chacune spécifique d'un SNP, recouvertes de fragments d'ADN, appelés oligonucléotides. Ces oligonucléotides correspondent à la séquence située en amont du SNP. L'ADN, préalablement amplifié et segmenté, est ensuite mis en contact avec les microbilles et les fragments se fixent sur les oligonucléotides correspondants. La plaque est ensuite lavée afin d'éliminer les fragments d'ADN qui ne se sont pas fixés. Des nucléotides associés à des fluorochromes, rouge ou vert selon la base, sont ensuite déposés sur la plaque et vont s'additionner à l'oligonucléotide de manière complémentaire à l'ADN génomique, la couleur rouge ou vert permettra ainsi d'indiquer l'allèle observé. La plaque est scannée pour analyser la fluorescence et déterminer les génotypes des SNPs de la puce.

Depuis 10 ans, le nombre de SNPs caractérisés sur les puces ADN a été multiplié et il est désormais possible d'analyser jusqu'à 2,5 millions de polymorphismes (SNPs et quelques CNVs) par puce. De même, une même plaque peut servir au génotypage de plusieurs individus simultanément, réduisant ainsi le coût des procédures expérimentales.

Le développement des projets HapMap [12] et 1000 Génomes [4] a largement contribué à l'augmentation de la couverture des puces. En effet, bien qu'utilisant des stratégies différentes, les deux grandes sociétés de génotypage, Affymetrix et Illumina, ont utilisé les SNPs identifiables à l'aide du projet HapMap. Affymetrix a initialement basé le choix des SNPs présents sur les puces en fonction de leur position sur le génome et de leur fréquence dans la population d'intérêt pour essayer d'assurer une couverture homogène du génome alors que la stratégie développée par Illumina utilise les tagSNPs permettant de capturer le plus de variabilité génétique possible.

4.2 Bases de données bioinformatiques

La première GWAS aboutissant à des associations significatives a été publiée en 2005 [50]. Depuis, le nombre de GWAS publiées chaque année n'a cessé d'augmenter et on recense aujourd'hui plus de 1 900 études publiées [40]. Cette croissance rapide s'explique notamment par l'émergence de plusieurs bases de données utilisées pour réaliser des GWAS.

4.2.1 dbSNP

La base de données « SNP database », connue sous le nom de dbSNP, est une base de données publique qui a pour but de répertorier les variations génétiques recensées chez plusieurs espèces animales dont l'être humain [5]. Les SNPs ne sont pas les seuls polymorphismes à être renseignés dans cette base, les indels ou les microsatellites y sont également inclus. Au total, plus de 60 millions de polymorphismes sont actuellement décrits dans cette banque. Cette base de données représente une source d'information majeure pour la communauté scientifique puisque le but de ce catalogue est d'annoter de façon précise les polymorphismes découverts dans le génome. Ainsi, lorsqu'un SNP est répertorié dans la base de données, des informations concernant sa localisation, ses conséquences fonctionnelles possibles ou encore les fréquences alléliques sont disponibles. De plus, lorsque des publications relatives à un SNP existent, la fiche dbSNP de ce site les recense et propose de liens vers celles-ci. Par conséquent, cette base de données constitue un outil majeur dans les études génomiques en permettant de connaître rapidement et facilement les caractéristiques d'un polymorphisme.

4.2.2 Le projet HapMap

Les haplotypes constituent une information essentielle dans les études génétiques. Ainsi, en 2002, un consortium international a entrepris une cartographie des haplotypes dans plusieurs populations : le projet HapMap [12, 51]. Ce projet s'est déroulé en trois phases.

Lors de la phase I du projet, environ 1 million de SNPs ont été génotypés chez 270 individus originaires de 4 populations différentes :

- 90 nigériens (Yoruba d'Ibadan) formant 30 trios (deux parents et un enfant)
- 90 résidents des Etats-Unis originaires d'Europe du Nord et d'Europe de l'Ouest formant 30 trios
- 45 japonais non apparentés de la région de Tokyo

- 45 chinois non apparentés de la région de Pékin

La phase II du projet a eu pour but d'augmenter le nombre de SNPs en génotypant environ 3,1 millions de SNPs chez les mêmes individus [52].

Enfin, la phase III visait à augmenter la densité de SNPs génotypés et le nombre d'individus inclus dont certains issus de populations non étudiées dans les phases précédentes [53] :

- Pour les populations originelles du projet, 90 nouveaux individus africains, 90 nouveaux individus d'origine européenne et 45 individus pour chacune des deux populations asiatiques ont été rajoutés au projet. Environ 5 millions de polymorphismes, dont des CNVs, ont été génotypés pour ces individus.
- Sept nouvelles populations (des afro-américains, des chinois résidant aux Etats-Unis, des mexicains résidant aux Etats-Unis, des toscans, des kenyans de la tribu Luhya, des kenyans masaïs et des indiens Gujarati vivant aux Etats-Unis) ont été ajoutées au projet. Environ 1,4 millions de SNPs ont été génotypés pour ces individus.

L'intéret du projet HapMap est multiple.

Tout d'abord, les données obtenues par le projet, et en particulier les génotypes, ont été mises gratuitement à la disposition de la communauté scientifique constituant une source d'information importante. De plus, il a permis d'identifier et de caractériser une grande partie des SNPs fréquents (dont la fréquence de l'allèle mineur est supérieure à 5%) dans plusieurs populations. Ainsi, le projet permet de connaître la fréquence de ces SNPs au sein des différentes populations génotypées. Enfin, l'inclusion de trios dans certaines populations a permis une cartographie des haplotypes. Etant donné l'existence de blocs haplotypiques au sein du génome (voir Introduction, chapitre 2.3), beaucoup des SNPs génotypés dans ce projet sont corrélés entre eux. La cartographie des haplotypes a alors permis la caractérisation des déséquilibres de liaison pour les paires de SNPs d'une même région. Cela a conduit à l'identification d'un nombre plus restreint de SNPs représentant la même information génétique qu'avec l'ensemble des polymorphismes. De tels SNPs sont appelés tagSNP (Figure 16). On estime ainsi que seulement 600 000 SNPs permettraient de capturer la diversité des 10 millions de SNPs fréquents du génome. Cette découverte a largement été utilisée pour la conception des puces de génotypage (voir Introduction, chapitre 4.1).

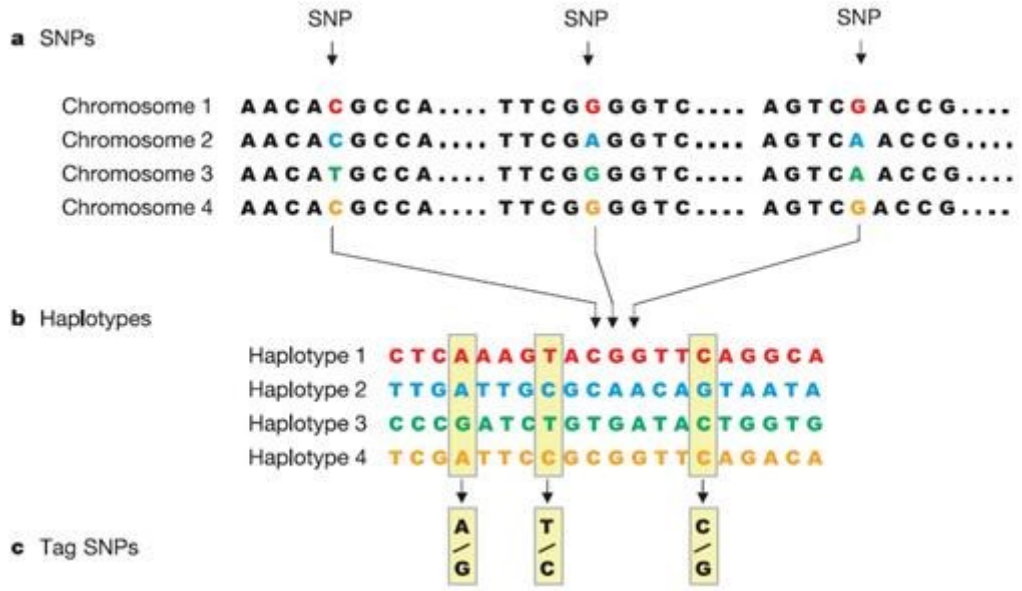


Figure 16. Notion de "tagSNP". (a) Identification de 3 SNPs (en couleur) dans une portion chromosomique. (b) Reconstruction des haplotypes constitués de 20 SNPs dont les trois identifiés précédemment. (c) Détermination de 3 tagSNPs dont la connaissance suffit pour identifier les 4 haplotypes de la population. Par exemple, un profil G-T-C pour ces 3 tagSNPs correspond toujours à l'haplotype 3. (Tirée de [12])

Ce projet a ainsi grandement contribué au développement des études génétiques et génomiques [54].

4.2.3 Le projet 1000 Genomes

Plus récemment, le projet 1000 Genomes dont l'objectif est de séquencer le génome de 2500 individus issus de 28 populations différentes a vu le jour [4]. Le séquençage intégral de ces génomes vise principalement à identifier l'ensemble des polymorphismes et en particulier les SNPs de faible fréquence qui ne sont pas inclus dans le projet HapMap. A l'instar de ce dernier, le projet 1000 Genomes s'est déroulé en plusieurs phases se différenciant principalement par le nombre de polymorphismes inclus dans l'étude. Actuellement, 2 577 individus ont été séquencés et environ 80 millions de polymorphismes ont été identifiés. Cependant, certaines précautions doivent être prises concernant les données mises à disposition. En effet, en plus des nombreuses difficultés inhérentes au séquençage, la couverture moyenne est de 4x (un locus est « lu » 4 fois en moyenne), ce qui entraîne une certaine incertitude quant aux polymorphismes repérés. Pour la même raison, les génotypes ne sont donnés que sous la forme de probabilité. Enfin, la faible fréquence de certains

polymorphismes peut être à l'origine d'incertitude concernant leur qualité. En conséquence, les données mises à disposition sont régulièrement corrigées ou mises à jour et il est alors préférable d'utiliser des versions moins complètes mais plus fiables du projet.

Cependant, ce projet ouvre de larges perspectives concernant l'implication de polymorphismes peu fréquents dans les maladies multifactorielles. En effet, le rôle de ces mutations est aujourd'hui largement soutenu [55]. Enfin, les données de 1000 Genomes servent également de panel de référence pour de nombreuses problématiques telles que le phasage ou l'imputation.

4.3 Reconstruction des haplotypes

Considérons par exemple deux SNPs S_1 et S_2 dont les allèles respectifs sont A/a et B/b. Pour chacun de ces deux SNPs, trois génotypes sont possibles (AA, Aa ou aa pour S_1) et les techniques de génotypage donnent cette information pour chaque individu. Toutefois, ces données ne permettent pas de connaître la phase de ces allèles, c'est-à-dire les haplotypes (voir Introduction, chapitre 2.3). En effet, supposons que les génotypes d'un individu pour S_1 et S_2 soient respectivement Aa et Bb. Dans ce cas, deux paires d'haplotypes (AB/ab ou Ab/aB) peuvent être observés chez cet individu. Les méthodes d'haplotypage visent à attribuer une paire d'haplotypes à chaque individu d'un groupe à partir des génotypes de ce groupe (Figure 17).

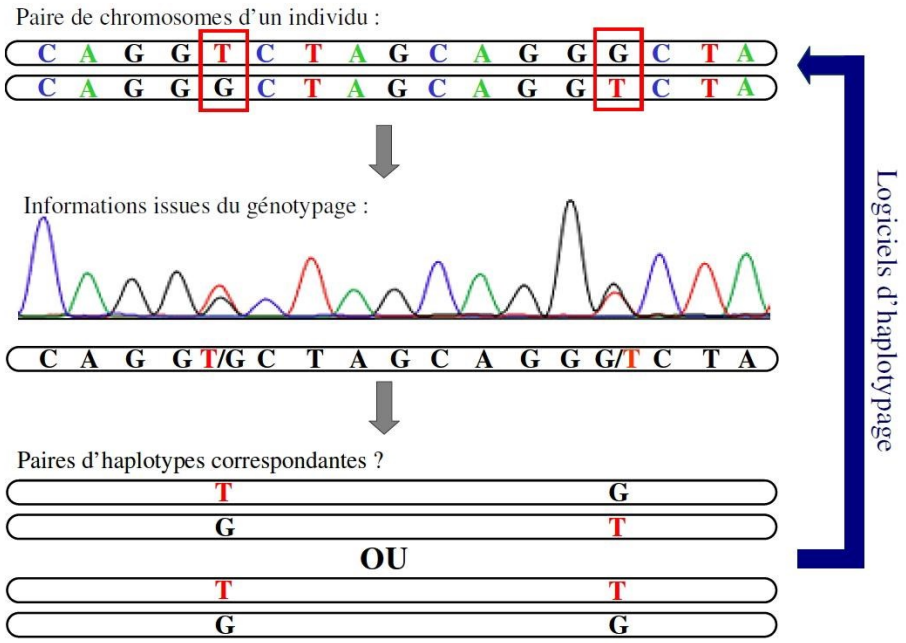


Figure 17. Problématique de l'haplotypage

Plusieurs méthodes ont été développées afin de reconstruire les haplotypes. Les méthodes combinatoires étudient tous les haplotypes possibles et n'en conservent qu'une paire pour chaque individu en se basant sur un critère de parcimonie [56] ou de phylogénie [57, 58]. Des méthodes d'inférence statistique ont été développées et se basent sur la vraisemblance en recherchant les haplotypes qui sont les plus probables étant donnés les génotypes observés [59, 60]. Enfin, plus récemment, des méthodes bayésiennes utilisant des a priori biologiques et des modèles de coalescence ont été mises au point [61-63]. Ces dernières méthodes semblent être les plus précises [64].

De nombreuses revues détaillent ces différentes méthodes [65-67].

4.4 Imputation

L'imputation vise à déterminer les génotypes de SNPs absents de la puce en se basant sur les informations de déséquilibre de liaison et des taux de recombinaison des haplotypes mesurés à l'aide d'un panel de référence (Figure 18). Ces panels de référence sont constitués d'un grand nombre de données haplotypées renseignant un nombre de SNPs supérieur à celui de marqueurs présents sur la puce. Généralement, les données de grands projets comme 1000 Genomes ou Hapmap sont utilisées.

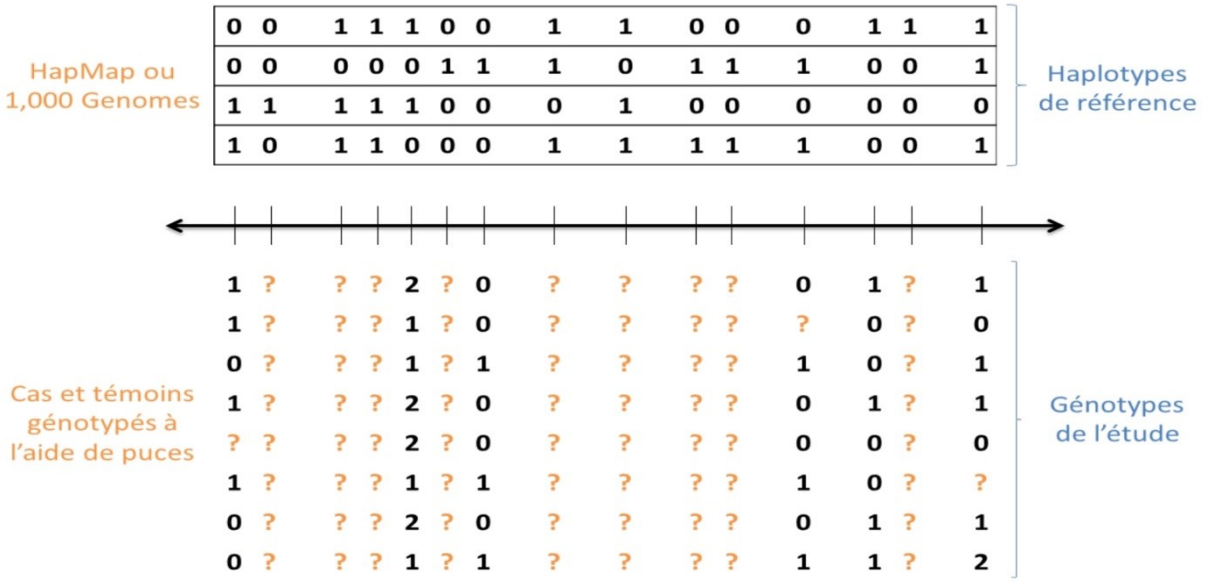


Figure 18. Représentation schématique du problème d'imputation. Les génotypes manquants de la puce seront imputés à l'aide des haplotypes d'un panel de référence. (Adaptée de [68])

Les puces de génotypage sont conçues à partir d'une liste de polymorphismes déterminée a priori permettant de capturer la majorité de l'information génétique. Bien que leur

développement permette le génotypage d'un nombre important de SNPs (jusqu'à 2,5 millions), il reste un grand nombre de polymorphismes non renseignés. Additionnellement, les SNPs génotypés d'une puce à l'autre ne sont pas identiques, ce qui rend souvent difficile la réplication des résultats obtenus d'une GWAS à l'autre. Enfin, certains SNPs peuvent être exclus d'une étude GWAS lors des contrôles de qualité en raison d'un trop grand nombre de données manquantes.

Les méthodes d'imputation développées au cours de ces dernières années peuvent apporter une réponse à ces problématiques en permettant :

- la connaissance de génotypes de SNPs absents de la puce de génotypage
- la caractérisation des données manquantes
- la comparaison de jeux de données obtenus par l'intermédiaire de puces différentes
- l'augmentation de la puissance statistique de l'étude
- la caractérisation plus fine de régions chromosomiques identifiées lors d'une GWAS

En pratique, la qualité de l'imputation dépend grandement de la taille du panel de référence. Plus ce panel inclut de SNPs et d'individus, plus la caractérisation des déséquilibres de liaison est fine et plus l'inférence des SNPs est fiable. Cependant, la complexité de l'imputation et les temps de calcul augmentent avec la taille du panel. Pour cette raison, il peut être très efficace de réaliser l'imputation à partir de données phasées car cette étape de reconstruction des haplotypes permet un gain de temps considérable tout en assurant une qualité d'imputation quasi-identique [69].

5 Recherche d'association dans les études génome-entier

5.1 Contrôle qualité

Cette étape est primordiale lors de la réalisation d'une GWAS puisque des données de mauvaise qualité peuvent empêcher l'identification de résultats ou la découverte de « faux » résultats. Ce contrôle qualité doit être effectué sur les données de génotypage qui sont ensuite elles même utilisées pour vérifier la qualité de la cohorte.

5.1.1 Génotypage

Le génotypage est un procédé expérimental, bien qu'automatisé, et de nombreux biais peuvent alors perturber cette étape. Une fois le génotypage effectué, plusieurs vérifications indispensables doivent être effectuées afin de s'assurer d'une bonne qualité des données obtenues.

Dans un premier temps, dès le génotypage terminé, l'analyse de la fluorescence de la plaque est effectuée à l'aide d'un logiciel développé par l'entreprise fabriquant la puce. Une première phase de contrôle est réalisée à ce moment-là pour vérifier que l'expérimentation s'est bien déroulée et que les génotypes peuvent être obtenus de façon fiable. Les SNPs dont les données ne sont pas fiables sont alors exclus de l'analyse. De la même façon, les SNPs pour lesquels une portion trop importante d'individus n'a pas été génotypée sont également exclus.

Il est également important de noter que tous les polymorphismes génotypés ne sont pas utilisables dans une GWAS. Par exemple, certaines puces renseignent des CNVs mais les méthodes d'analyse de ces polymorphismes diffèrent de celles utilisées pour les SNPs. Les CNVs seront donc mis à part et exclus de la GWAS afin d'être étudiés séparément.

Dans un second temps, les SNPs dont la fréquence de l'allèle mineur (MAF pour Minor Allele Frequency en anglais) est inférieure à un certain seuil, généralement 1%, sont exclus de l'étude. En effet, l'analyse statistique des SNPs de faible fréquence est problématique dans une GWAS. De plus, les puces visent à génotyper les polymorphismes fréquents dans la population étudiée. Ainsi, un SNP dont la MAF est trop faible ne correspond pas aux résultats

attendus et la fiabilité du génotypage de ce SNP est alors remise en cause. En pratique, si un tel phénomène est observé, il est néanmoins possible de vérifier si l'algorithme utilisé pour le génotypage a « failli », si la fluorescence est de mauvaise qualité ou si cela est véritablement spécifique à notre population.

Enfin, un test est réalisé pour chacun des SNPs pour s'assurer qu'ils vérifient l'équilibre d'Hardy-Weinberg [70]. Cet équilibre est un modèle central de la génétique des populations (voir Introduction, chapitre 2.1). S'il n'est pas respecté, en plus de raisons biologiques (consanguinité, sélection, migration), cela peut être dû à un problème lors du génotypage et à une répartition non conforme des génotypes. Les SNPs dont la p-valeur est inférieure à un certain seuil sont exclus de l'analyse. Dans le cas d'étude du type « cas-témoins », ces tests sont effectués au sein de la population de témoins puisqu'une déviation chez les cas peut être la conséquence d'une association à la pathologie.

5.1.2 Cohorte

Une fois le contrôle qualité du génotypage effectué, il est important de s'assurer de la bonne qualité de la cohorte étudiée.

De la même façon que cela est fait pour le génotypage, les individus pour lesquels une fraction trop importante de SNPs n'a pu être génotypée sont exclus de l'analyse.

Les études d'association se focalisent sur des individus supposés non apparentés. En pratique, il est possible que des erreurs de manipulation (étiquetage des tubes d'ADN, patients génotypés plusieurs fois…) surviennent ou que deux individus apparentés soient inclus dans l'étude. Afin de vérifier que les individus de la cohorte sont effectivement indépendants, une estimation de la proportion du génome dite « Identical By Descent » (IBD), c'est-à-dire héritée d'un même ancêtre commun, est calculée pour chaque paire d'individus. Si cette proportion est trop importante, les individus sont considérés comme apparentés et un individu de chaque paire concernée (en général celui avec le plus de données manquantes) sera exclu de l'étude.

Enfin, une structure sous-jacente dans la population peut mener à la découverte de fausses associations génétiques (voir Introduction, chapitre 3.2.1.2). Il est alors fondamental d'utiliser les données de génotypage de la cohorte afin de détecter si certains individus sont issus d'une population génétiquement différente de celle de la majorité. Les individus atypiques sont

généralement exclus de l'analyse. Deux méthodes sont principalement employées pour détecter la stratification au sein d'une cohorte.

La première méthode est implémentée dans le logiciel STRUCTURE [71]. Cette méthode utilise des méthodes bayésiennes afin d'inférer le nombre de sous-populations qui composent la cohorte et l'appartenance de chaque individu à ces sous-populations. Les probabilités d'appartenance peuvent ensuite être représentées sur un graphique afin de déterminer visuellement les individus s'éloignant de leur population d'origine (Figure 19).

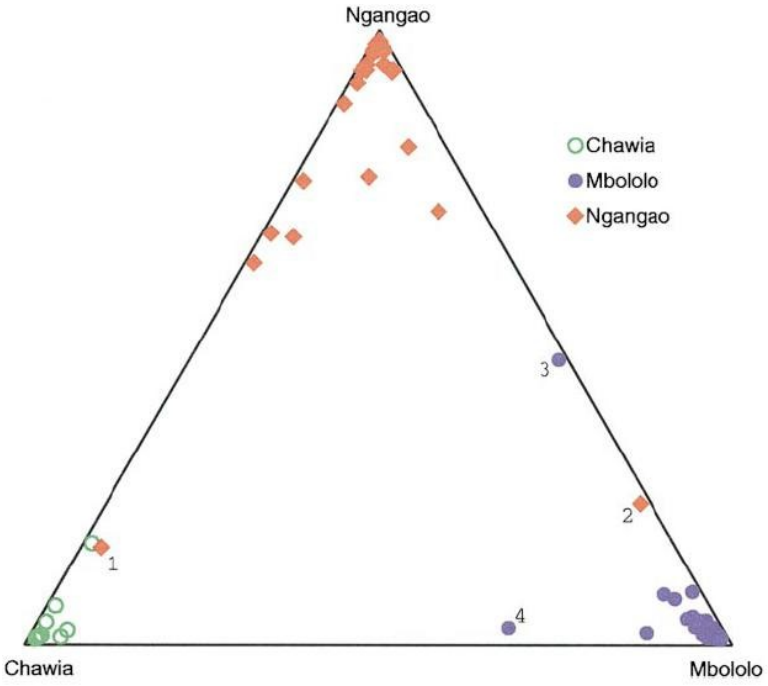


Figure 19. Exemple de stratification. Les données sont les génotypes d'une population menacée d'oiseaux T. Helleri. Chaque point représente la probabilité moyenne d'appartenance d'un individu à une des trois populations. Les meilleurs résultats ont été obtenus pour 3 populations. Les populations d'origine de chaque individu ont été connues après l'obtention des résultats. Les points 1-4 représentent des individus potentiellement outliers étant donné qu'ils s'éloignent de leur population d'origine. (Tirée de [71])

Afin d'améliorer les résultats de cette méthode, la connaissance de la population d'origine de chaque individu peut être prise en compte dans le modèle. De plus, il est possible avec cette méthode d'assumer qu'un patient appartient à une unique population ou alors qu'il est issu de plusieurs population distinctes. Les résultats de cette analyse permettent donc d'éliminer les sujets qui s'éloignent de la cohorte. Néanmoins, cette méthode n'est pas adaptée à des jeux de données importants.

Plus récemment, une méthode d'Analyse en Composantes Principales (ACP), implémentée dans le logiciel EIGENSTRAT, a été proposée pour déterminer la structure d'une cohorte [72]. Les données d'entrée sont les génotypes des individus. L'ACP va déterminer des combinaisons indépendantes de SNPs, appelés vecteurs propres, qui capturent la plus grande partie de la variabilité génétique. Ainsi, le premier vecteur propre explique une plus grande partie de la variabilité observée au sein de la population que le deuxième vecteur propre (indépendant du premier), et ainsi de suite. Ces vecteurs propres constituent alors un nouveau repère dans lequel chaque individu peut être représenté. Les individus s'éloignant de manière trop importante du nuage de points correspondant à sa population d'origine est alors considéré comme un outlier (Figure 20).

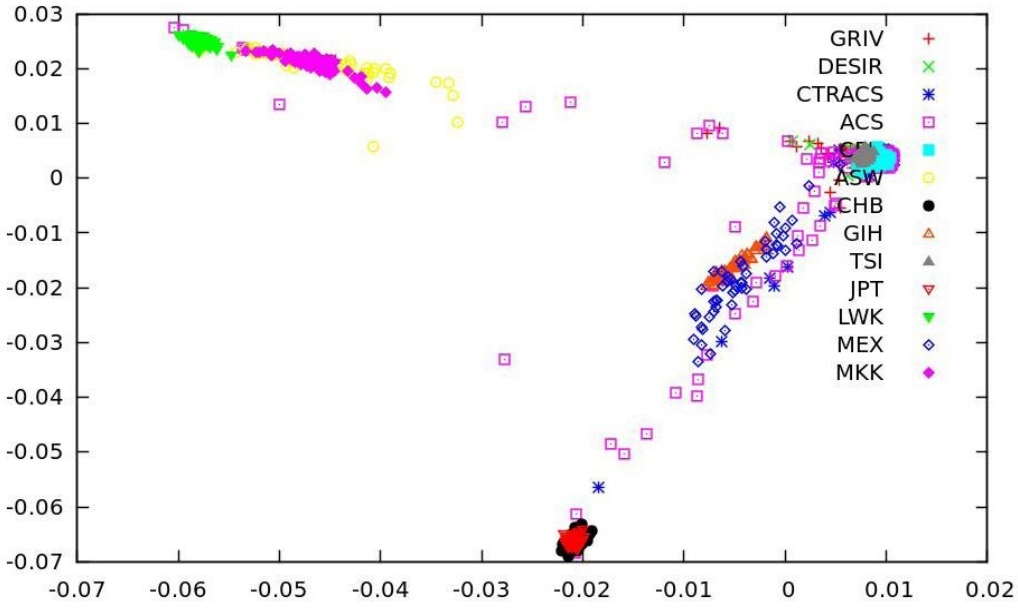


Figure 20. Représentation des deux premiers axes de l'analyse en composantes principales réalisée à partir de plusieurs dizaines de milliers de SNPs avec le logiciel EIGENSTRAT. Chaque individu est représenté par un point. Les populations du projet HapMap3 sont bien différenciées. Certains individus de cohortes analysées au sein du laboratoire (GRIV, DESIR, ACS et CTRACS) s'éloignent du nuage d'individus de leur population d'origine et sont donc considérés comme outliers.

Cette dernière méthode est désormais privilégiée pour différentes raisons. D'une part, les temps de calcul nécessaires à la réalisation de l'ACP sont plus faibles que pour la méthode précédente et restent raisonnables pour de gros jeux de données. D'autre part, les individus aberrants repérés par STRUCTURE le sont aussi par l'ACP mais cette dernière permet une description plus fine, puisque continue, de la structure. Ainsi, les résultats obtenus par cette

méthode peuvent être intégrés comme facteurs de confusion dans les études afin de corriger la possible stratification de la population considérée.

5.2 Facteurs de confusion

Les facteurs de confusion sont également une source potentielle de biais lors d'une GWAS. Un facteur de confusion est une variable associée au trait étudié et qui peut par conséquent influer sur le résultat d'association avec le marqueur étudié (Figure 21).

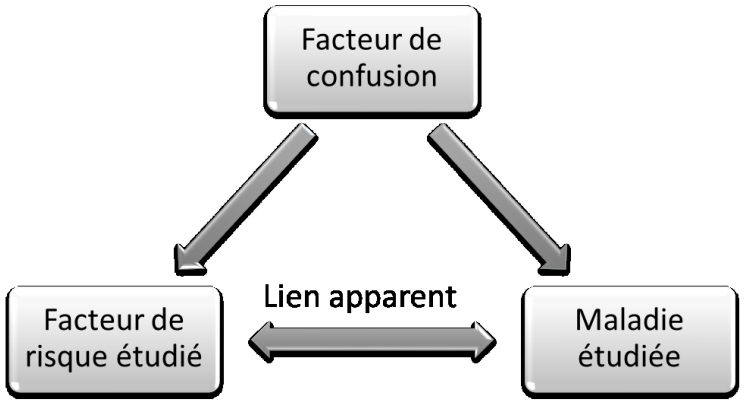


Figure 21. Problématique des facteurs de confusion. Le lien entre le facteur de confusion et à la fois le facteur de risque et la maladie étudiés conduisent à l'apparition d'un lien entre le facteur de risque et la maladie.

Les facteurs environnementaux ou comportementaux peuvent être des facteurs de confusion. Par exemple, l'âge, le sexe, l'alimentation ou la consommation de tabac sont des facteurs associés au cancer et peuvent donc biaiser les tests d'association entre un SNP et le développement d'un cancer s'ils ne sont pas pris en compte. Ces variables auront d'autant plus d'effet que leur distribution au sein de la cohorte est inégale. Si lors d'une étude cas-contrôle sur le cancer, la proportion de fumeurs est plus importante chez les cas, alors des marqueurs pourront être associés au phénotype alors qu'en réalité ils sont associés au tabagisme. Lors d'une GWAS, la stratification est un des principaux facteurs de confusion. De même, les mesures du phénotype à l'aide de différentes méthodes ou dans des centres différents constituent de fréquentes sources de biais.

Ces variables doivent être prises en compte lors des analyses d'associations. Les études bibliographiques permettent de déterminer les facteurs de confusions potentiels. De plus, lors de la constitution d'une cohorte, le phénotype n'est pas la seule variable mesurée chez les individus. En effet, les facteurs de confusion potentiels sont pris en compte. Afin de

déterminer les facteurs à inclure dans les analyses statistiques, il est fréquent de mesurer les corrélations entre ces variables et le phénotype étudié. Toutes les variables significativement corrélées au trait étudié sont incluses en tant que « covariables » dans l'étude.

5.3 Mesure statistique de l'association entre le polymorphisme et le phénotype

Une fois les différentes étapes du contrôle qualité effectuées et les covariables déterminées, l'objectif est de déterminer l'ensemble des polymorphismes associés de façon significative au phénotype étudié. Ces recherches d'association sont effectuées pour chacun des SNPs à l'aide de méthodes statistiques et notamment les tests statistiques d'hypothèses et les méthodes de régression.

5.3.1 Tests statistiques d'hypothèses

Le but d'un test d'hypothèses est de déterminer statistiquement si une hypothèse par défaut, dite nulle, peut être rejetée ou non au profit d'une hypothèse alternative au regard des données à disposition. Par exemple, il est possible d'effectuer un test d'hypothèses pour s'assurer que l'âge moyen des cas est égal à celui de la population contrôle ou que les fréquences alléliques d'un polymorphisme sont identiques au sein de ces deux populations.

Considérons un phénomène aléatoire modélisé par une variable aléatoire dont la distribution est connue à un paramètre $\theta \in \Theta$ près. Dans le cadre d'un problème de test, on cherche, en se basant sur les données à disposition, à confronter l'hypothèse selon laquelle ce paramètre appartient à un sous-ensemble Θ_0 de Θ à celle selon laquelle le paramètre appartient à un autre sous-ensemble disjoint du premier. Un test statistique constitue alors un critère de décision entre ces deux hypothèses.

5.3.1.1 Formulation des hypothèses

Lors d'un test statistique, deux hypothèses sont testées : l'hypothèse nulle et l'hypothèse alternative.

Faire une hypothèse notée (H_0) sur le paramètre θ consiste à se donner un sous-ensemble Θ_0 de l'espace des paramètres Θ. L'hypothèse, appelée hypothèse nulle, s'écrit :

$$(H_0) : \theta \in \Theta_0$$

L'origine de cette dénomination se trouve dans les situations où l'absence d'un certain effet est recherchée.

Une hypothèse alternative, notée (H_1), est une hypothèse :

$$(H_1) : \theta \in \Theta_1 \text{ telle que } \Theta_1 \subset \Theta \text{ et } \Theta_0 \cap \Theta_1 = \emptyset$$

L'hypothèse alternative consiste alors à se donner un sous-ensemble Θ_1 de Θ disjoint de Θ_0.

Par exemple, pour s'assurer que la taille moyenne θ au sein d'un échantillon est égale à la taille moyenne m au sein de la population générale, il est possible de tester les hypothèses suivantes :

$$(H_0) : \theta = m \text{ contre } (H_1) : \theta \neq m$$

Un test statistique revient, à partir des données observées, à prendre une des deux décisions suivantes : accepter l'hypothèse nulle (H_0) (et donc décider que $\theta \in \Theta_0$), ou rejeter (H_0) au profit de l'hypothèse alternative (H_1) (et donc décider que $\theta \in \Theta_1$). Un test statistique teste toujours une hypothèse nulle contre une hypothèse alternative.

Dans le cas de la recherche d'association entre un SNP et un phénotype, les hypothèses testées sont :

- (H_0) : pas d'association entre le SNP et le phénotype
- (H_1) : le SNP est associé au génotype

En théorie des tests, les deux hypothèses n'ont pas un rôle symétrique puisque l'hypothèse nulle est privilégiée dans le sens où elle est conservée sauf si les données observées conduisent à la rejeter. L'analogie entre un test statistique et un procès d'assise illustre bien ce concept : lors d'un procès, le suspect est considéré comme innocent tant que la preuve de sa culpabilité n'est pas apportée. De la même façon, l'hypothèse nulle est considérée comme vraie tant que les données ne permettent pas de la rejeter. Ainsi, accepter (H_0) correspond au verdict « acquitter faute de preuve ». Mathématiquement, il est préférable de dire dans ce cas que (H_0) n'est pas rejetée. Si les données permettent de rejeter l'hypothèse nulle, celle-ci est rejetée au profit de l'hypothèse alternative formulée. En effet, pour une même hypothèse nulle, les conclusions d'un test statistique peuvent varier en fonction de l'hypothèse alternative.

5.3.1.2 Région d'acceptation et zone de rejet

Une fois les hypothèses à tester clairement établies, la procédure de décision à partir d'un test statistique se base sur les données observées. Un test statistique peut alors être représenté comme une fonction de ces données.

Lors de la confrontation entre (H_0) et (H_1), un test pur est une statistique Φ à valeurs dans {0,1}, associée à la stratégie suivante : pour les données observées (notées x), (H_0) est acceptée si $\Phi(x) = 0$ et rejetée au profit de (H_1) si $\Phi(x) = 1$. La région de rejet du test est donc l'ensemble $\Phi^{-1}(1)$ des données possibles telles que $\Phi(x) = 1$. La région de rejet, et la fonction Φ, caractérisent entièrement un test pur. Un tel test fournit alors un critère de décision binaire.

Dans le cas classique où le paramètre ne peut avoir qu'une valeur ($\Theta_0 = \{\theta_0\}$), afin de déterminer ces zones, une statistique de test *T(x)* calculée à partir des données est évaluée. Cette statistique *T(x)* est une réalisation d'une variable aléatoire *T* dont la distribution est connue si l'hypothèse nulle est vraie. La zone de rejet peut alors être caractérisée par l'ensemble des données possibles pour lesquelles *T(x)* vérifie certaines conditions relatives à la distribution de *T*. Par exemple, les données pour lesquelles *t* est supérieure à un certain seuil (test unilatéral) ou pour lesquelles *T(x)* est supérieure à un certain seuil ou inférieure à un autre seuil (test bilatéral) constituent la zone de rejet (Figure 22). La fonction de test Φ peut alors s'écrire :

- $\Phi(x) = \mathbb{I}_{T(x) \geq s}$ dans le cas d'un test unilatéral
- $\Phi(x) = \mathbb{I}_{T(x) \leq s_1} + \mathbb{I}_{T(x) \geq s_2}$ dans le cas d'un test bilatéral

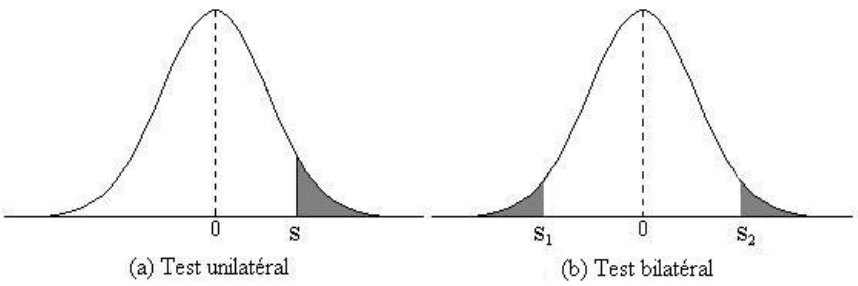


Figure 22. Exemples de zones de rejet dans le cas où la statistique de test suit une loi normale $\mathcal{N}(0,1)$. (a) Dans le cas d'un test unilatéral, la zone de rejet correspond aux données pour lesquelles la statistique de test est supérieure à s. (b) Dans le cas d'un test bilatéral, la zone de rejet correspond aux données pour lesquelles la statistiques de test est inférieure à s_1 ou supérieure à s_2

Le calcul de la statistique de test permet alors, en comparant la valeur obtenue avec le(s) seuil(s), de prendre une décision relative aux hypothèses testées. Cependant, cette prise de décision implique un risque d'erreur.

5.3.1.3 Erreurs de décision

Lors de la prise de décision consécutive à un test statistique, il est possible de faire deux types d'erreurs. Dans le cas où l'hypothèse nulle est rejetée alors qu'elle est vraie en réalité, l'erreur est dite de première espèce. Si l'hypothèse nulle n'est pas rejetée alors qu'elle est fausse en réalité, l'erreur est dite de deuxième espèce (Tableau 1).

		Décision	
		(H_0)	(H_1)
Réalité	(H_0)	Décision correcte	Erreur de deuxième espèce
	(H_1)	Erreur de première espèce	Décision correcte

Tableau 1. Les deux types d'erreurs possibles lors d'un test d'hypothèses. Une erreur de première espèce est commise lorsque l'hypothèse nulle est rejetée à tort et l'erreur de deuxième espèce est réalisée lorsque l'hypothèse n'est pas rejetée alors que l'hypothèse alternative est vraie.

A chaque type d'erreur correspond un risque et une probabilité de la commettre.

5.3.1.3.1 Risque de première espèce, niveau et probabilité critique

L'erreur de première espèce d'un test Φ est l'application, notée α, qui à chaque $\theta_0 \in \Theta_0$ donne la probabilité de prendre la mauvaise décision et de rejeter l'hypothèse nulle à tort :

$$\alpha : \Theta_0 \rightarrow [0,1]$$
$$\theta_0 \mapsto \alpha(\theta_0) = \mathbb{E}_{\theta_0}(\Phi) = P_{\theta_0}(x, \Phi(x) = 1)$$

Cette quantité est mesurée en utilisant la loi de probabilité associée au paramètre θ_0.

Un test Φ est de niveau ou de seuil α, avec $\alpha \in [0,1]$, si la probabilité maximale de commettre une erreur de première espèce est respectivement égale ou inférieure à α :

$$\sup_{\theta_0 \in \Theta_0} \alpha(\theta_0) = \alpha$$

Dans le cas où Φ peut s'écrire sous la forme $\Phi(x) = \mathbb{I}_{T(x) \geq s}$, avec s la constante pour laquelle Φ est de niveau α, la probabilité critique, ou p-valeur (p-value en anglais), est définie comme la probabilité d'observer une statistique de test supérieure à celle obtenue avec les données x :

$$p(x) = \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T \geq T(x))$$

De même, si $\Phi(x) = \mathbb{I}_{T(x) \leq s_1} + \mathbb{I}_{T(x) \geq s_2}$, la p-valeur est définie par :

$$p(x) = \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(\{T \geq T(x)\} \cup \{T \leq T(x)\})$$

Si $\Theta_0 = \{\theta_0\}$, alors rejeter l'hypothèse nulle est équivalent à observer une p-valeur inférieure au niveau α.

En pratique, le seuil est déterminé arbitrairement par la personne qui réalise le test. Classiquement, ce seuil est fixé à 10%, 5%, 1% ou 0,1%. Par exemple, un seuil à 5% revient à dire que la personne qui réalise le test s'accorde au plus 5% de chance de rejeter à tort une hypothèse nulle.

5.3.1.3.2 Risque de deuxième espèce et puissance statistique

Le risque de deuxième espèce d'un test Φ est la fonction, notée β, qui à chaque $\theta_1 \in \Theta_1$ associe la probabilité de ne pas rejeter l'hypothèse nulle :

$$\begin{aligned} \beta : \Theta_1 &\to [0,1] \\ \theta_1 &\mapsto \beta(\theta_1) = P_{\theta_1}(x, \Phi(x) = 0) \end{aligned}$$

La fonction puissance du test est l'application notée γ qui à chaque $\theta_1 \in \Theta_1$ associe la probabilité de rejeter l'hypothèse nulle :

$$\begin{aligned} \gamma : \Theta_1 &\to [0,1] \\ \theta_1 &\mapsto 1 - \beta(\theta_1) = P_{\theta_1}(x, \Phi(x) = 1) \end{aligned}$$

La puissance d'un test peut alors être vue comme la capacité du test à rejeter l'hypothèse nulle à raison. Il est donc souhaitable de réaliser des tests dont la puissance est la plus grande possible.

La puissance d'un test dépend de plusieurs facteurs comme par exemple la taille de l'échantillon ou l'écart réel du paramètre par rapport au paramètre attendu sous (H_0). En prenant l'exemple d'un test permettant de détecter l'association entre un polymorphisme et un phénotype, la puissance dépend de :

- la force de l'association entre ce polymorphisme et le phénotype : plus celle-ci est faible, moins le test est puissant.
- les fréquences alléliques du polymorphisme : plus la MAF du polymorphisme est faible, moins le test a de puissance.
- la taille de l'échantillon : la puissance d'un test croît avec le nombre de patients inclus dans l'étude

Il est également important de noter que diminuer le risque de première espèce d'un test se fait au détriment de l'erreur de deuxième espèce. Choisir de diminuer le niveau d'un test revient à augmenter l'erreur de deuxième espèce de celui-ci. Il convient en conséquence de s'orienter vers un compromis entre ces deux types d'erreurs, c'est-à-dire de s'accorder un risque faible de commettre une erreur de première espèce tout en minimisant le risque de ne pas par rejeter une hypothèse nulle alors que celle-ci est fausse.

Quelle que soit la méthode statistique utilisée pour mesurer l'association d'un polymorphisme avec une maladie, un test statistique est réalisé pour évaluer sa significativité statistique mesurée par une p-valeur.

5.3.2 Correction des tests multiples

5.3.2.1 Problématique

Dans le cadre de comparaisons multiples comme pour les GWAS, le risque de première espèce augmente avec le nombre de tests réalisés et n'est plus contrôlé. En d'autres termes, en se fixant un risque global de première espèce au niveau α, retenir la liste de tests dont les p-valeurs individuelles sont inférieures au seuil α ne permet pas de s'assurer que le risque de rejeter à tort au moins une hypothèse nulle est inférieur à ce niveau. Pire, la probabilité de commettre une telle erreur augmente considérablement.

Considérons la réalisation de *m* tests. De même que pour la réalisation d'un seul test, les prises de décision peuvent conduire à des erreurs (Tableau 2).

		Décision		
		Significatif	Non significatif	Total
Réalité	Hypothèse nulle	V	U	m_0
	Hypothèse alternative	S	T	$m - m_0$
	Total	R	m - R	m

Tableau 2. Classification des résultats suite aux tests de m hypothèses.

La probabilité d'obtenir au moins un résultat significatif, soit S > 0, est donée par la formule suivante :

$$P(S > 0) = 1 - P(S = 0) = 1 - (1 - \alpha)^m$$

La probabilité d'obtenir un résultat significatif augmente donc avec le nombre de tests réalisés et cet événement devient quasi certain lorsqu'un très grand nombre de tests (Figure 23).

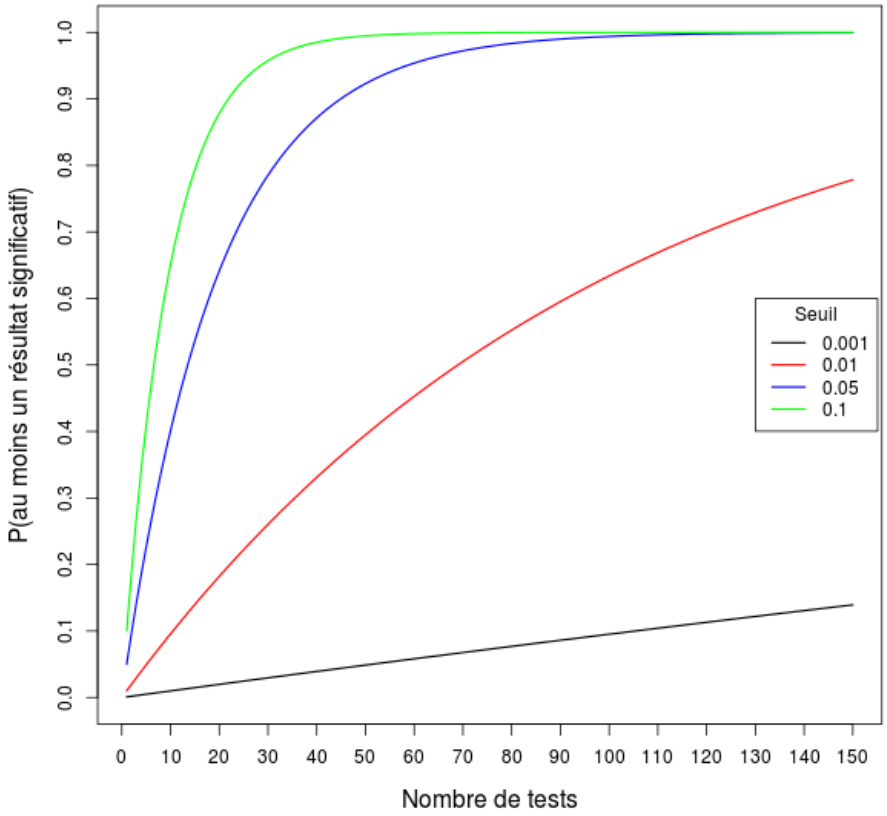


Figure 23. Probabilité d'obtenir au moins un résultat significatif en fonction du nombre de tests réalisés pour différents seuils. Quel que soit le seuil, cette probabilité augmente avec le nombre de tests effectués. De même, plus le seuil est grand, plus la probabilité augmente rapidement.

Afin de contrôler le risque d'erreur de première espèce à un seuil α, une première approche consiste à s'assurer que la probabilité de rejeter à tort au moins une hypothèse nulle, appelée Family-Wise Error Rate (FWER), est inférieure à α :

$$FWER = P(V > 0) \leq \alpha$$

La méthode la plus utilisée pour contrôler ce FWER est la méthode de Bonferroni [73].

5.3.2.2 Correction de Bonferroni

La grande utilisation de la correction de Bonferroni s'explique en grande partie par sa simplicité. En effet, pour chaque test *i* des *m* réalisés, la p-valeur ajustée p_i^{adj} s'exprime en fonction de la p-valeur d'origine *p* :

$$p_i^{adj} = \min(mp, 1)$$

Cela équivaut à comparer la p-valeur d'origine à un nouveau seuil α_{adj}, appelé seuil de Bonferroni, défini comme le seuil fixé divisé par le nombre de tests effectués :

$$\alpha_{adj} = \frac{\alpha}{m}$$

Cette modification permet bien de contrôler le risque de première espèce au seuil fixé α.

$$FWER = P(V > 0) = P(\bigcup_{I_0} \{p_i^{adj} \leq \alpha\}) \leq \sum_{i=1}^{m_0} P(p_i^{adj} \leq \alpha) \leq \sum_{i=1}^{m_0} \frac{\alpha}{m} \leq \frac{m_0 \alpha}{m} \leq \alpha$$

L'obtention de SNPs dont les p-valeurs sont inférieures au seuil de Bonferroni constitue le meilleur résultat possible, puisque cette méthode assure une probabilité de commettre au moins une erreur inférieure au seuil fixé α. Cette correction facile à mettre en œuvre présente cependant un inconvénient majeur dans la mesure où elle est très conservatrice et certaines hypothèses nulles ne sont pas rejetées alors qu'elles ne sont pas vraies en réalité (faux négatifs, T > 0). De plus, la méthode de Bonferroni se base sur l'hypothèse d'indépendance des tests réalisés alors que celle-ci n'est pas respectée dans les GWAS en raison des déséquilibres gamétiques existant entre les SNPs.

Une approche alternative, pour s'affranchir de ces problèmes, consiste à estimer le taux de fausses découvertes (FDR, False Discovery Rate en anglais) [74, 75].

5.3.2.3 Taux de Fausses Découvertes (FDR)

Le FDR est une approche différente visant à estimer le nombre de faux positifs parmi les hypothèses rejetées pour un seuil donné [74, 75] :

$$FDR = \begin{cases} E\left[\frac{V}{R}\right], & si\ R\ > 0 \\ 0, & si\ R = 0 \end{cases}$$

Le FDR peut être calculé pour n'importe quel seuil t de p-valeur. Soient $0 \leq t \leq 1$ et $p_1, p_2, \ldots, p_m$ les m p-valeurs triées par ordre croissant,

$$FDR(t) = E\left[\frac{V(t)}{R(t)}\right] \approx \frac{E[V(t)]}{E[R(t)]}$$

où F(t) désigne le nombre de vraies hypothèses nulles dont la p-valeur est inférieure à t et S(t) désigne le nombre de p-valeurs inférieures à t. L'approximation du FDR(t) par le ratio n'est valable que si un grand nombre m d'hypothèses sont testées, ce qui est le cas pour une GWAS. En reprenant les notations précédentes (Tableau 2), il est possible d'estimer le nombre attendu de faux positifs :

$$E[F(t)] = m_0 \times t$$

Or, m_0 n'est pas connu, mais il est possible d'estimer la proportion $\hat{\pi}_0 = \frac{m_0}{m}$. Une fois ce paramètre estimé, il est possible de calculer FDR(t) :

$$FDR(t) = \frac{\hat{\pi}_0\, m\, t}{S(t)}$$

Cette estimation du FDR étant réalisable pour tout seuil t, il semble donc naturel de la calculer en utilisant l'ensemble des p-valeurs des m tests effectués comme seuil. Cela aboutit à la détermination pour chaque p-valeur p_i d'une valeur appelée q-valeur $q(p_i)$:

$$q(p_i) = \min_{t \geq p_i} FDR(t)$$

L'avantage principal de cette q-valeur est sa facilité d'interprétation : la q-valeur d'un test représente la proportion estimée de faux positifs si ce test est considéré comme significatif. En pratique, un seuil de q-valeur est fixé, généralement à 25% et l'ensemble des tests dont la q-valeur est inférieure à ce seuil est considéré comme significatif.

De nombreuses adaptations, comme le FDR local par exemple, et méthodes d'estimations ont été proposées pour cette correction [76-78].

5.3.3 Analyse d'un phénotype quantitatif

Dans certaines études, le phénotype analysé est décrit par une variable quantitative. Les associations avec de tels phénotypes sont mesurées à l'aide d'une régression linéaire. Les méthodes de régression visent à modéliser la relation entre une variable dite à expliquer et un ensemble de variables dites explicatives.

Notons Y la variable à expliquer et $X = (X_0, X_1, \ldots, X_p)$ l'ensemble des variables explicatives. Les techniques de régression ont pour but de déterminer une fonction f des variables explicatives estimant au mieux la variable à expliquer Y :

$$Y = f(X) + \varepsilon$$

où f est une fonction à déterminer et ε représente des termes d'erreurs. Généralement, lors d'une GWAS, les modèles de régression linéaire sont privilégiés et ainsi l'ensemble des fonctions f étudiées sont des fonctions affines :

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

avec $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$ le vecteur de paramètres à estimer. Cette estimation est réalisée à partir d'observations $y = (y_1, \ldots, y_n)$ et $x_i = (x_{i1}, \ldots, x_{ip})$ chez n individus. Ainsi, le modèle de régression conduit aux équations suivantes :

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases}$$

soit,

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = X\beta + \varepsilon$$

La méthode des moindres carrés ordinaires permet l'estimation du vecteur β des coefficients :

$$\beta = (X'X)^{-1}X'Y$$

avec X' désignant la matrice transposée de X.

Une fois ces coefficients estimés, un test d'hypothèses est réalisé pour chacun d'entre eux afin de confronter les hypothèses suivantes :

$$(H_0) : \beta_j = 0 \text{ contre } (H_1) : \beta_j \neq 0$$

En d'autres termes, les hypothèses (H_0) et (H_1) stipulent respectivement l'absence ou non d'effet de la variable explicative *j* sur le phénotype étudié. Ce test aboutit à une p-valeur qui comparée au seuil fixé pour le test permet de conclure quant à la significativité statistique de l'influence de la variable sur le phénotype étudié. Naturellement, l'attention est portée sur le coefficient relatif au génotype du SNP dont l'association avec le phénotype est étudiée.

Cependant, certains phénotypes quantitatifs introduisent une notion de durée et mesurent le temps écoulé avant la survenue d'un événement comme le décès [79] ou la progression vers le sida [80]. Dans de tels cas, la régression linéaire multiple n'est pas adaptée et l'analyse de ces phénotypes s'effectue à l'aide d'une régression de Cox [81]. Ces types de régressions vont modéliser le risque de survenue de l'événement considéré en fonction du temps et des différentes variables explicatives.

5.3.4 Analyse d'un phénotype qualitatif

Dans le cadre d'une étude d'association génome entier comparant une population de cas à une population témoin, le phénotype étudié est l'appartenance à la population des cas ou des témoins. Plusieurs méthodes, comme des tests d'indépendance ou des méthodes de régression, peuvent être utilisées afin de mesurer l'association entre un polymorphisme et un tel phénotype.

5.3.4.1 Test d'indépendance du Khi-deux

Considérons, par exemple, un SNP A dont les allèles sont a_1 et a_2 pour lequel on étudie l'association avec un phénotype sous le modèle dominant pour l'allèle a_1. L'échantillon est composé de N individus, répartis en N_C cas et N_T témoins. Au sein de cet échantillon, N_{a_1} individus sont homozygotes pour l'allèle a_1, N_{a_2} individus sont homozygotes pour l'allèle a_2 et N_{het} sont hétérozygotes.

Les individus inclus dans l'étude peuvent être distingués selon deux critères dans une table de contingence (Tableau 3) : l'appartenance à l'une des deux populations et la présence ou non de l'allèle a_1 dans leur génotype.

	Cas	Témoins	Total
Porteurs de l'allèle a_1	$N_{a_1}^C + N_{het}^C$	$N_{a_1}^T + N_{het}^T$	$N_{a_1} + N_{\text{het}}$
Non -porteurs de l'allèle a_1	$N_{a_2}^C$	$N_{a_2}^T$	N_{a_2}
Total	N_C	N_T	N

Tableau 3. Table de contingence d'une étude d'association de type « cas/témoins » en considérant l'allèle a_1 dominant

L'objectif est de déterminer s'il existe une association significative entre le fait de porter ou non l'allèle a_1 et l'appartenance à une des deux populations. Le test statistique permettant de prendre une décision est un test d'indépendance du χ^2 entre ces deux variables dont les hypothèses sont :

- (H_0) : les deux variables sont indépendantes (pas d'association)
- (H_1) : les deux variables ne sont pas indépendantes (association)

L'idée du test est de comparer les effectifs observés dans le tableau de contingence aux effectifs attendus si les deux variables sont indépendantes. Dans un premier temps, les effectifs attendus sont calculés comme le produit des effectifs marginaux divisé par l'effectif total. Par exemple, le nombre théorique $E_{a_1}^C$ de cas porteurs de l'allèle a_1 est :

$$E_{a_1}^C = \frac{N_c \times (N_{a_1} + N_{\text{het}})}{N}$$

La statistique de test utilisée pour prendre la décision est la somme des différences entre les effectifs observés et théoriques divisées par l'effectif théorique :

$$T = \sum_{i,j} \frac{(O_i^j - E_i^j)^2}{E_i^j}$$

où O_i^j désigne l'effectif observé pour la modalité *i* de la première variable (ici porteur ou non de l'allèle a_1) et la modalité *j* de la seconde variable (ici cas ou témoins).

Si les variables sont indépendantes ((H_0) est vraie), T suit une loi du χ^2 à (i-1)x(j-1) degrés de liberté. Dans l'exemple, T suit une loi du χ^2 à 1 degré de liberté.

Ensuite, le seuil α du test est fixé (généralement à 0,05) et les zones d'acceptation et de rejet du test peuvent alors être déterminées (Figure 24).

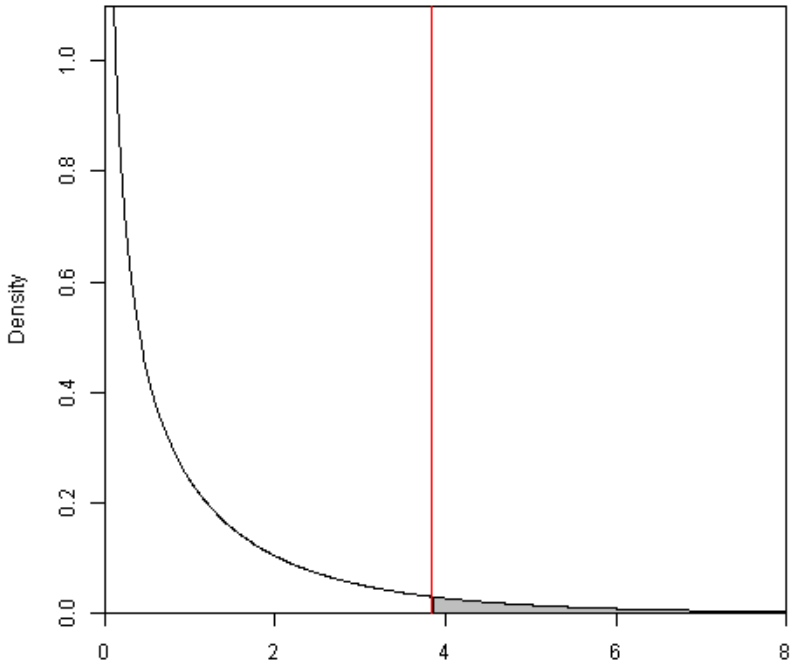


Figure 24. Zones d'acceptation et de rejet (gris) d'un test d'indépendance du Khi-deux à un degré de liberté pour un seuil de 0,05. Toute statistique de test supérieure au 95ème centile de la distribution (3,84 ; ligne rouge) conduit au rejet de l'hypothèse nulle au profit de l'hypothèse alternative.

Finalement, la statistique de test *T* est comparée au quantile 1- α de la distribution du Khi-deux à un degré de liberté. Si T est supérieur à ce quantile, l'hypothèse d'indépendance des deux variables est rejetée au profit de l'hypothèse alternative (dépendance des variables). Dans ce cas, il est possible de conclure à une association entre le fait de porter l'allèle a_1 et

l'appartenance à une des deux populations. Dans le cas contraire, la décision prise à l'aide du test est l'absence d'association.

Toutefois, l'association n'est pas totalement caractérisée par la simple connaissance de la p-valeur du test. En effet, le fait de porter l'allèle a_1 augmente-t-il ou diminue-t-il la probabilité d'appartenir à la population des cas ? Pour répondre à cette question, un rapport des chances, ou odds ratio (OR) en anglais, est estimé à partir du tableau de contingence :

$$OR = \frac{(N_{a_1}^{C} + N_{het}^{C}) \times N_{a_2}^{T}}{(N_{a_1}^{T} + N_{het}^{T}) \times N_{a_2}^{C}}$$

L'odds ratio varie entre 0 et +∞. Un odds ratio supérieur à 1 indique que le fait de porter l'allèle a_1 augmente le risque d'appartenir à la population cas. Il est à noter que le risque croît avec la valeur de l'odds ratio. Inversement, ce risque est diminué si l'OR est inférieur à 1, et plus la valeur de l'OR sera proche de 0, plus la diminution du risque sera importante. De plus, il est possible de calculer un intervalle de confiance à 100 x (1 – α) % qui peut être vu schématiquement comme un intervalle auquel le véritable OR appartient avec une probabilité (1 – α) :

$$IC_{1-\alpha} = \left[e^{\ln(OR) - Z_{1-\alpha}SE(\ln(OR))};\ e^{\ln(OR) + Z_{1-\alpha}SE(\ln(OR))}\right]$$

où SE(ln(OR)) = $\sqrt{\frac{1}{N_{a_1}^{C}+N_{het}^{C}} + \frac{1}{N_{a_1}^{T}+N_{het}^{T}} + \frac{1}{N_{a_2}^{C}} + \frac{1}{N_{a_2}^{T}}}$ et $Z_{1-\alpha}$ est le centile $1-\alpha$ de la loi normale $\mathcal{N}(0,1)$.

Il est utile de noter que les deux variables ne sont pas indépendantes pour un test de seuil α si la valeur 1 est exclue de cet intervalle.

Ce type d'analyse présente l'avantage d'être rapide et facile à mettre en œuvre mais elle ne permet pas de prendre en compte les facteurs de confusion pouvant influer sur l'appartenance à l'une des deux populations.

Afin de pallier cet inconvénient, des tests d'indépendance stratifiés par des covariables qualitatives, les tests de Cochran-Mantel-Haenszel, ont été développés [82]. Schématiquement, une table de contingence est créée pour chacune des modalités de la variable selon laquelle la stratification est faite, puis les résultats de chaque strate sont

compilés pour obtenir le résultat global. Toutefois, l'application de cette méthode reste limitée en raison de ses inconvénients. En plus de l'augmentation du nombre de tables de contingences à analyser et des problèmes d'effectif pouvant survenir, la stratification n'est possible que par des variables qualitatives. Par conséquent, des variables quantitatives, comme l'âge par exemple, doivent nécessairement être découpées en classes afin d'être intégrées à l'analyse. Une telle transformation implique nécessairement une perte d'information et peut également poser des problèmes de justification du découpage effectué. Afin d'outrepasser ces désavantages, les modèles de régression ont été privilégiés.

5.3.4.2 Régression logistique

La régression logistique est utilisée pour la modélisation de la relation entre une variable binomiale (dont les deux valeurs possibles sont 0 ou 1) et des variables explicatives. Classiquement, ce genre de régression est utilisé pour modéliser la présence (cas) ou non (témoins) d'une pathologie.

La régression logistique vise à modéliser la probabilité $\pi(x) = \mathrm{P}(\mathrm{Y} = 1 \mid \mathrm{X} = \mathrm{x})$. Cependant, la régression linéaire multiple n'est pas adaptée dans ce cas puisque cette dernière peut aboutir à des valeurs qui ne sont pas interprétables en tant que probabilité. La régression est alors réalisée sur une transformation de la quantité π à l'aide d'une fonction de lien. La fonction de lien la plus fréquemment utilisée est la fonction logit définie par :

$$\begin{aligned} logit : \mathbb{R} &\to [0,1] \\ \mathrm{x} &\mapsto ln\left(\frac{x}{1-x}\right) \end{aligned}$$

Ainsi, le modèle utilisé dans une régression logistique s'écrit :

$$logit(\pi) = X\beta$$

Le vecteur β est alors estimé à l'aide du maximum de vraisemblance et un test statistique confrontant les mêmes hypothèses que précédemment permet de conclure sur l'impact ou non de la variable explicative.

Il existe d'autres fonctions de lien pouvant être utilisées pour les régressions sur des variables binomiales [83].

5.3.5 Qualité des résultats

Outre la détection des SNPs significativement associés au phénotype considéré, l'analyse des résultats inclut le contrôle de la bonne qualité des résultats. Pour cela, la distribution des p-valeurs obtenues est comparée à celle attendue si tous les SNPs ne sont pas associés au caractère. En effet, si tous les SNPs sont indépendants du phénotype (toutes les hypothèses (H_0) sont vraies), la distribution des p-valeurs suit une loi uniforme sur [0,1]. Etant donné le faible nombre attendu de SNPs associés au phénotype, la distribution de p-valeurs obtenus est globalement celle d'une loi uniforme sur [0,1].

La comparaison s'effectue sur les quantiles des deux distributions. Un p-ième q-quantile d'une variable X est la valeur $x_{p/q}$ pour laquelle une certaine proportion ($100 \times \frac{p}{q}$ %) des observations constituant la distribution est inférieure à ce quantile :

$$P(X < x_{p/q}) = \frac{p}{q}$$

Le diagramme quantile-quantile permet une comparaison graphique de la distribution obtenue de p-valeurs à la loi uniforme attendue. Dans ce diagramme, les quantiles de la distribution attendue sont représentés en abscisse et les quantiles de la distribution observée en ordonnées (Figure 25). Si les deux distributions suivent la même loi de probabilité, le nuage de points suit la droite d'équation y= x étant donné que leurs quantiles sont identiques. Toute déviation précoce de la distribution observée indique un biais dans l'analyse. Celui-ci peut-être dû à l'omission d'un facteur de confusion dans l'analyse ou alors à une stratification au sein de l'échantillon. Il convient donc de rechercher les causes de cette déviation afin d'effectuer l'analyse de nouveau pour obtenir des résultats non biaisés.

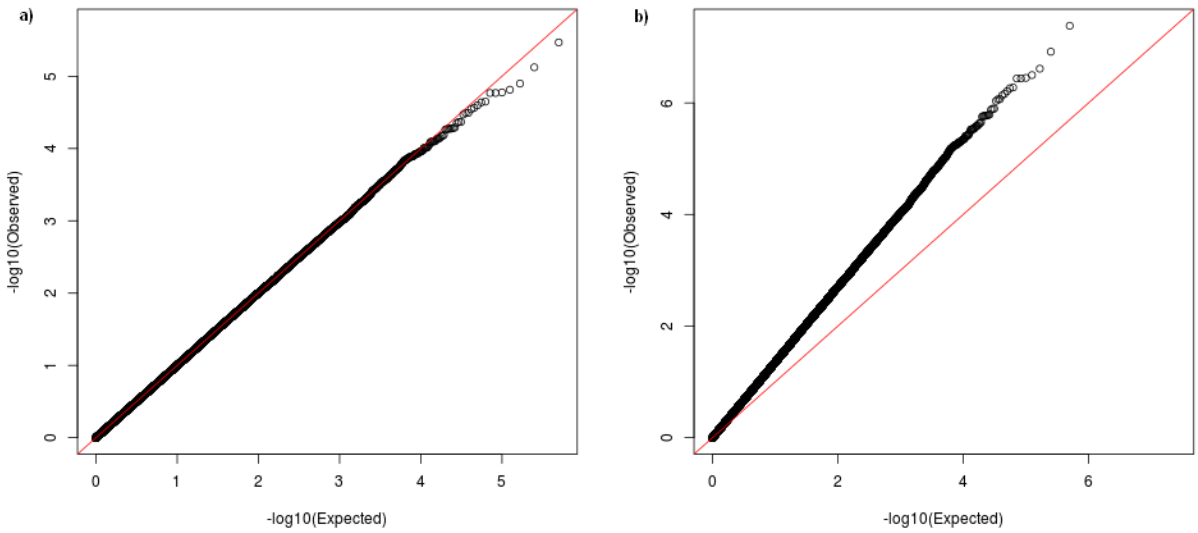


Figure 25. Exemples de diagrammes quantile-quantile. a) Le nuage de points se superposent avec la droite d'équation y = x (en rouge) indiquant que les deux distributions sont identiques. b) Le nuage de points s'écartent de la droite y = x (en rouge) et les deux distributions ne sont pas identiques, ce qui témoigne de résultats peu fiables.

Après s'être assuré de l'absence de biais dans la mesure des associations, les p-valeurs sont représentées à l'aide d'un « manhattan plot ». Ce graphique représente l'opposé du logarithme des p-valeurs le long des chromosomes (Figure 26). Sur ce graphique apparaît le seuil de Bonferroni (ou le seuil de FDR) afin d'identifier rapidement les SNPs significativement associés au phénotype.

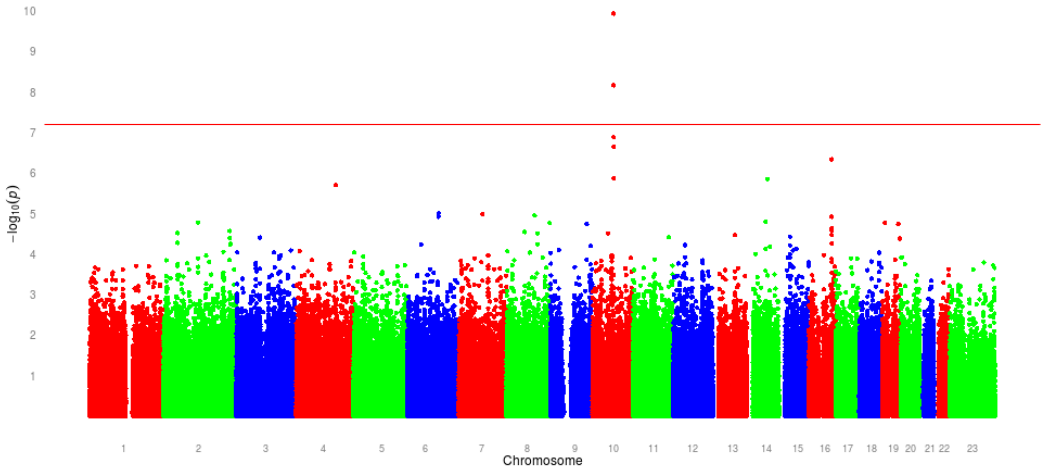


Figure 26. Exemple de Manhattan plot. Les p-valeurs sont représentées le long des chromosomes. Chaque point représente la p-valeur d'association d'un SNP avec le phénotype. Tous les points situés au-dessus du seuil de Bonferroni (ligne rouge) sont significativement associés au phénotype.

5.4 Analyses complémentaires

L'analyse des résultats d'une GWAS ne s'arrête pas une fois les signaux significatifs identifiés. En effet, plusieurs caractéristiques biologiques doivent être étudiées et il convient d'apporter les réponses à de nombreuses questions. Tout d'abord, quelle est la position des signaux identifiés ? Sur quels chromosomes sont-ils situés, se trouvent-ils dans un gène ou non, une région promotrice ? Cela permet de se faire une première idée des caractéristiques fonctionnelles. Si ces signaux sont situés dans un gène, il faut également savoir s'ils ont une conséquence sur la protéine : sont-ils des mutations synonymes, entraînent-ils des pertes ou des gains de fonction de la protéine ? De même, plusieurs bases de données [84-87] répertorient les influences des SNPs sur le niveau d'expression des gènes en testant si les différents génotypes d'un SNPs ont un niveau d'expression d'ARNm significativement différents les uns des autres. Cette recherche d'information permet de mieux appréhender l'impact biologique du polymorphisme. Néanmoins, il est important de garder à l'esprit que les signaux identifiés peuvent ne pas être les signaux causaux. Ainsi, il convient de rechercher tous les SNPs en déséquilibre de liaison avec les SNPs repérés et de répondre aux questions précédentes pour chacun d'entre eux.

De plus, si d'autres GWAS ont déjà été menées sur le même phénotype, il est important de vérifier la réplication des signaux identifiés. Même si cela n'était pas forcément facile à mettre en place en raison notamment des différentes techniques de génotypage utilisées dans chacune des études, l'imputation (voir Introduction, chapitre 4.4) permet de contourner ce problème. De même, les méta-analyses combinent différentes études indépendantes afin d'augmenter la puissance statistique de l'étude en accroissant le nombre de sujets de l'étude. Plusieurs méthodes ont été développées. Il est possible de n'utiliser que les p-valeurs des marqueurs ou d'utiliser le sens de l'association. La finalité de ces études est de combiner, pour chaque SNP, les associations des différentes études afin d'en obtenir une globale. De plus, ces études peuvent conduire à l'identification de nouveaux signaux qui n'ont pas été repérés dans les études individuelles. Avec la multiplication des GWAS, nombre d'études se focalisent sur le même phénotype rendant par conséquent possible la réalisation de nombreuses méta-analyses [88-92].

5.5 Nouvelles approches d'exploitation des données de GWAS

Depuis 2005, les GWAS ont identifié de nombreux polymorphismes impliqués dans plusieurs pathologies. Cependant, ce type d'étude, en raison du grand nombre de variants testés, détecte

principalement les polymorphismes ayant un impact important sur le fonctionnement des gènes, alors que d'autres SNPs situés dans des gènes importants pour le phénotype seront passés sous silence car leur impact sur le fonctionnement du gène reste modéré. Afin de contrecarrer ces limitations, des approches complémentaires ont été développées ces dernières années afin d'intégrer la connaissance biologique dans les analyses d'association. L'approche la plus connue est celle s'appuyant sur la connaissance des voies de signalisation ou "biological pathways". Une voie de signalisation est un ensemble de phénomènes moléculaires conduisant à une modification biologique. Les molécules impliquées sont de plusieurs types puisque peuvent être mis en jeu des protéines, des nutriments ou des acides gras entre autres. Par exemple, la glycogènese est l'ensemble des actions ou interactions cellulaires ayant pour finalité la synthèse de glucose. Etant donné l'existence de ces voies de signalisation, une partie de la communauté scientifique a développé des techniques permettant de tester l'impact non plus d'un polymorphisme sur le phénotype observé mais des SNPs associés aux pathways sur ce phénotype. Les SNPs associés aux pathways seront tout simplement les SNPs des gènes impliqués dans ce pathway. Cette alternative permet de prendre en compte la connaissance biologique et rend possible l'identification de signaux plus modérés non détectés à cause de la stringence de la correction des tests multiples mais ayant une réelle implication sur le phénotype (Figure 27).

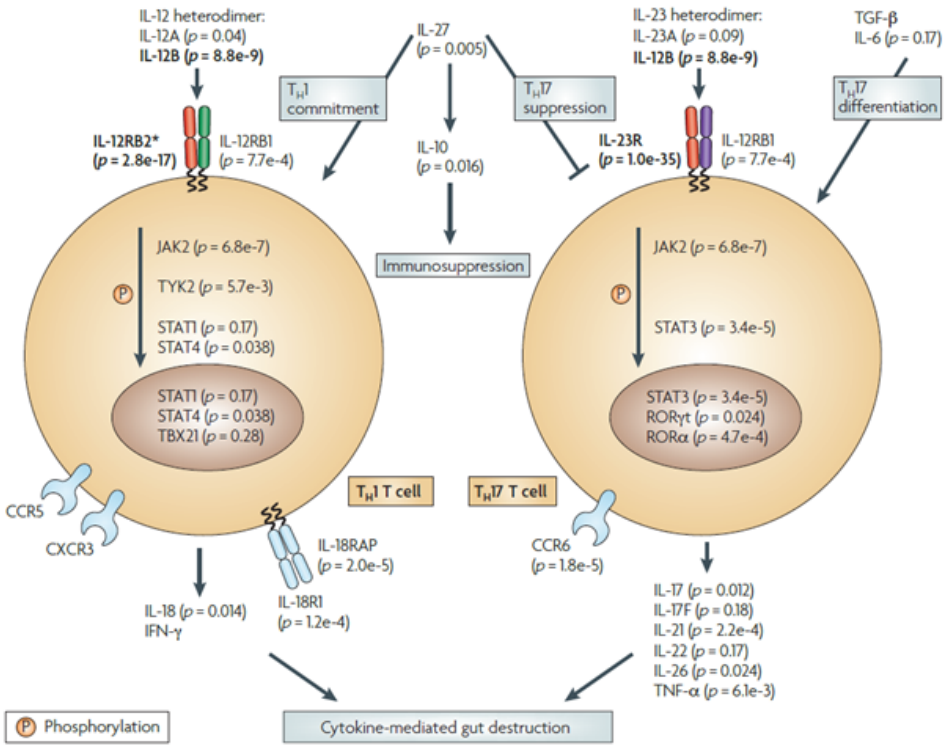


Figure 27. Exemple d'un pathway impliqué dans la maladie de Crohn.(Tirée de [93])

De plus, ces approches permettent d'avoir une meilleure compréhension des mécanismes sous-jacents en identifiant une fonction biologique particulière plutôt qu'un polymorphisme dont les répercussions biologiques peuvent être plus difficiles à appréhender. Enfin, l'implication d'un ensemble de gènes ou protéines offre potentiellement plus de pistes pour le développement de médicaments. Tous ces arguments expliquent l'engouement autour de ces méthodes et le nombre croissant d'études de ce type réalisées.

De nombreuses stratégies ont été mises en place afin de repérer le rôle de ces voies de signalisation dans les GWAS bien que de nombreuses limites comme l'influence du déséquilibre de liaison ou la redondance de pathways par exemple doivent être prises en compte. Plusieurs revues exposent les différentes problématiques inhérentes à ce type d'étude et les méthodes d'ores et déjà implémentées [93-97].

6 Peau et vieillissement

6.1 La peau, organe essentiel

La peau est l'organe le plus étendu et le plus lourd du corps humain. Cette enveloppe, dont le rôle majeur est de protéger l'individu, s'étend sur environ 2 m^2 et pèse entre 4,5 et 5 kilogrammes. Elle est composée de trois tissus superposés ayant chacun une fonction particulière : l'épiderme, le derme et l'hypoderme. Différents éléments, comme les poils et les glandes sébacés ou sudoripares, sont également présents dans la peau afin d'en assurer l'homéostasie (Figure 28).

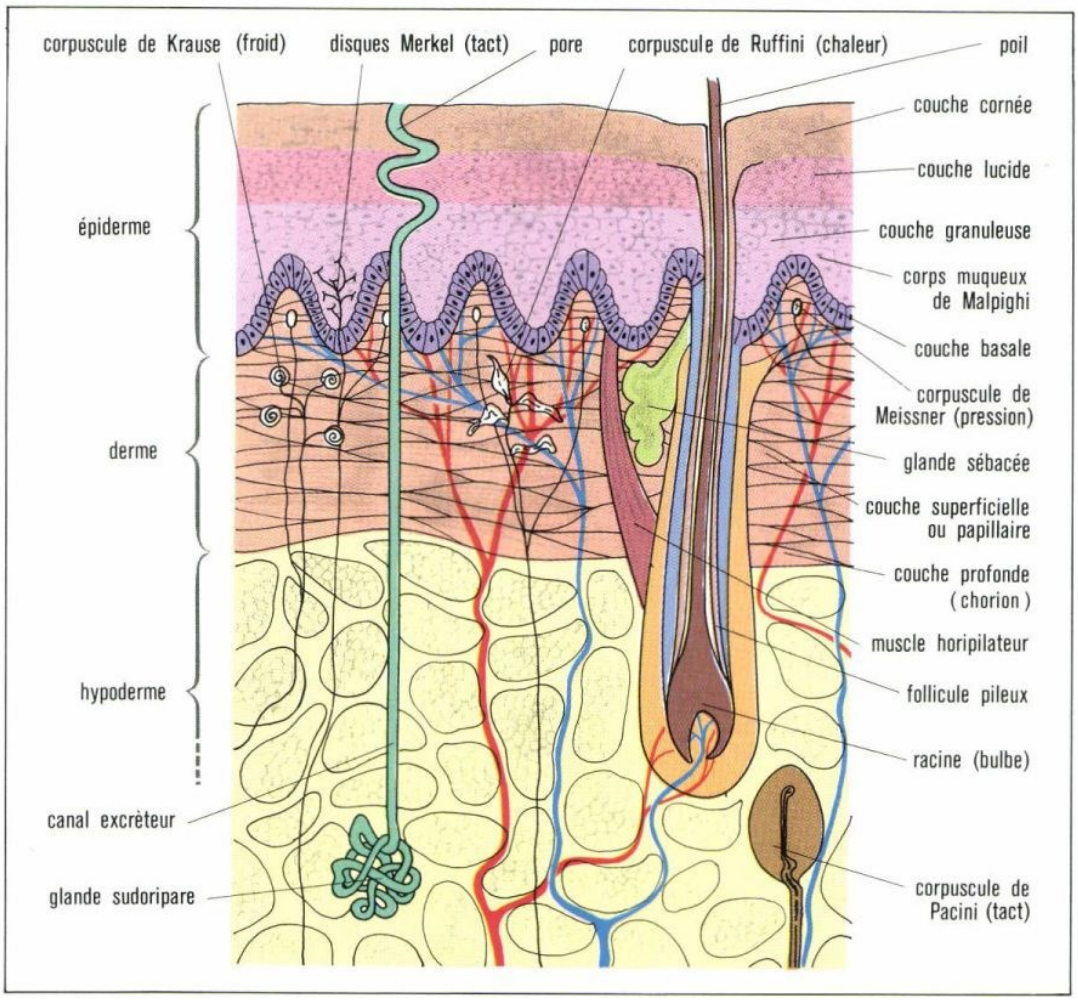


Figure 28. Représentation schématique d'une coupe histologique de la peau. (Tirée de http://cultureinvitrodepeau-tpe.e-monsite.com)

6.1.1 Epiderme

Selon la localisation sur le corps humain, l'épiderme est constitué de 4 à 5 couches superposées : de la couche cornée (couche supérieure) à la couche basale. Quatre grands types de cellules sont situés dans l'épiderme : les kératinocytes, les mélanocytes, les cellules de Langerhans et les cellules de Merkel (Figure 29).

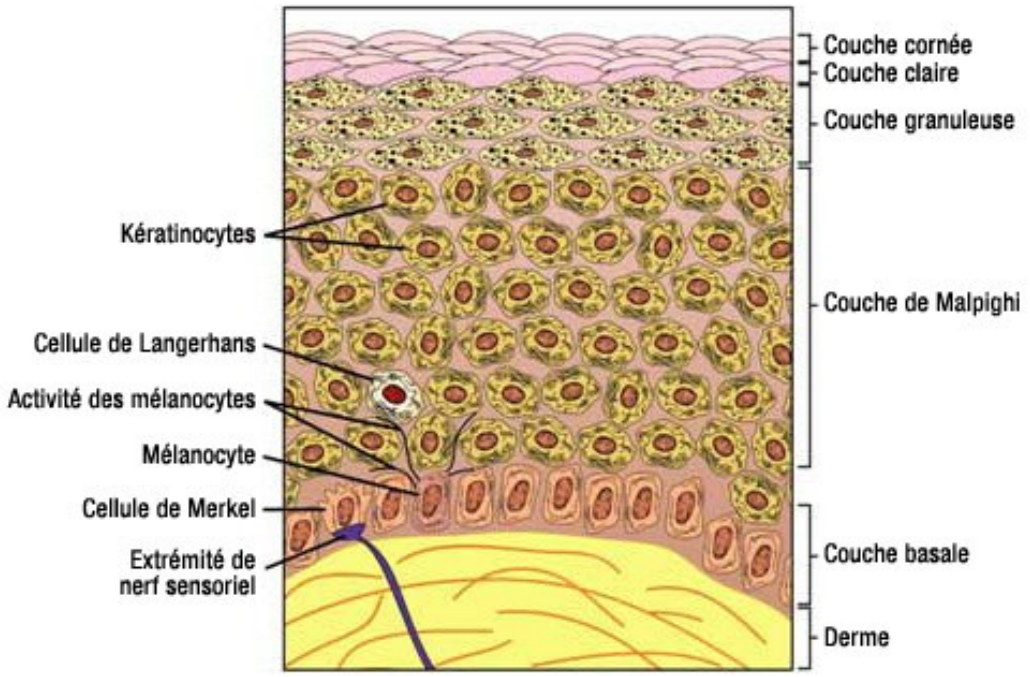


Figure 29.Représentation schématique d'une coupe transversale de l'épiderme. (Tirée de http://www.prevu.com)

Les kératinocytes sont les cellules majoritaires de cette couche de la peau et représentent 90% de l'épiderme. Ces cellules naissent dans la couche basale de l'épiderme puis, à la suite de plusieurs différentiations cellulaires, migrent vers la couche supérieure où elles perdent leur adhérence et tombent. La fonction principale de ces cellules est la protection contre les agressions environnementales telles que les éléments pathogènes ou les radiations d'ultraviolets (UV).

Les mélanocytes, qui ne représentent qu'1% des cellules de l'épiderme, ne sont présents que dans la couche basale. Ce sont de grandes cellules qui présentent de grands prolongements, appelés dendrites, entrant en contact avec les kératinocytes voisins. La fonction principale de ces cellules est la production de mélanine sous la forme de petits grains qui migrent dans les kératinocytes via les dendrites. La mélanine, dont la qualité et la quantité dépend des facteurs

génétiques de l'individu, détermine la couleur de la peau et assure un rôle protecteur contre les UV en préservant les noyaux des keratinocytes.

Les cellules de Langerhans sont des cellules dendritiques présentes dans toutes les couches de l'épiderme et constituent 2 à 7% des cellules de cette strate de la peau. Ces cellules détectent les corps étrangers ayant pénétré dans l'épiderme et migrent ensuite vers les ganglions lymphatiques du derme afin de les présenter aux lymphocytes. Ces cellules jouent donc un rôle fondamental dans l'immunité.

Enfin, les cellules de Merkel jouent un rôle mécano-récepteur et sont impliquées notamment dans la fonction du toucher. Elles sont situées dans la couche basale au contact des terminaisons nerveuses.

Outre les cellules, l'épiderme est composé également de lipides, à la fois des lipides de surface qui forment le film hydrolipidique et des lipides dits intra-épidermiques. Les glandes sébacées synthétisent les lipides de surface qui sont ensuite excrétés dans le sébum. Ces lipides se composent essentiellement de squalène (15%), de cire (25%) et de triglycérides (60%). Les lipides intra-épidermiques sont synthétisés par les kératinocytes. Ils sont composés de céramides (50%), de cholestérol (25%) et d'acides gras libres (25%).

L'épiderme a principalement une fonction de défense et de protection de l'organisme contre les stress mécaniques, les éléments pathogènes et les rayonnements UV. Enfin, en réaction à une exposition à certains rayons UV, les couches profondes de l'épiderme synthétisent la vitamine D3.

6.1.2 Derme

Le derme est un tissu conjonctif se divisant en deux parties : le derme papillaire ou superficiel et le derme réticulaire ou profond. Le derme est principalement constitué de fibroblastes qui ont pour rôle la synthèse de collagène, d'élastine, de la substance fondamentale et de glycoprotéines (Figure 30). Toutes ces protéines forment la matrice extracellulaire du derme permettant le maintien de la structure du tissu et l'adhérence des cellules. Le derme assure donc par l'intermédiaire de ce réseau de molécules le soutien de l'épiderme. De plus, cette couche de la peau est composée de cellules comme les leucocytes, les mastocytes et les macrophages impliqués dans le système immunitaire de l'organisme. Enfin le derme abrite de nombreux petits vaisseaux, des nerfs, des glandes sudoripares et sébacées. Cela lui confère

ainsi une fonction métabolique (apport d'oxygène, de nutriments,…), thermorégulatrice et sensorielle.

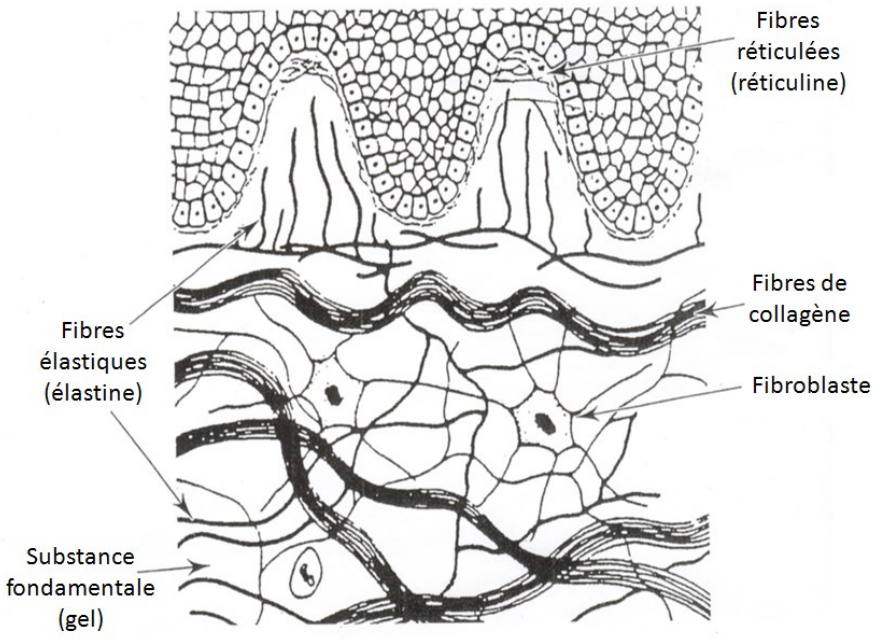


Figure 30.Représentation schématique d'une coupe transversale du derme. (Adaptée de http://cultureinvitrodepeau-tpe.e-monsite.com)

6.1.3 Hypoderme

L'hypoderme, également appelé tissu adipeux blanc est lié à la couche profonde du derme par des fibres de collagène et d'élastine formant des cloisons entre des boules graisseuses appelées lobules adipeux. Ces lobules constituent la majeure partie de l'hypoderme et sont des amas d'adipocytes. Les adipocytes sont des cellules dérivées de fibroblastes accumulant et stockant les graisses. La fonction de cette strate de la peau est dans un premier temps mécanique en amortissant les chocs et en luttant contre le froid mais l'hypoderme sert également de réserve énergétique.

6.2 Vieillissement général

Le vieillissement est caractérisé par la perte progressive de l'intégrité physiologique des tissus. Le passage du temps altère le fonctionnement de tous les organes et la capacité de renouvellement des cellules. Le vieillissement se traduit aussi bien d'un point de vue fonctionnel que psychomoteur. La majorité des fonctions du corps humain décline avec le passage du temps. Par exemple, la tension artérielle augmente, la respiration est moins

efficace, la solidité des os s'amoindrit, les déplacements sont moins aisés et les facultés sensorielles sont moins bonnes. Ces modifications se manifestent par des troubles visibles comme les tremblements, les troubles de l'équilibre ou encore des troubles cognitifs. Le vieillissement est le facteur de risque majeur de nombreuses maladies humaines telles que le cancer, le diabète ou les maladies cardio-vasculaires. Les modifications dues au passage du temps sont observables aussi bien du point de vue métabolique que cellulaire ou moléculaire.

6.2.1 Modifications métaboliques

Bien que les raisons n'en soient pas obligatoirement connues, de nombreuses fonctions métaboliques permettent de caractériser le processus du vieillissement comme la résistance à l'insuline, les modifications de composition du corps ou le déclin de certaines hormones par exemple. Une théorie du vieillissement relie directement le vieillissement au métabolisme en stipulant que plus le métabolisme sera important, plus le vieillissement le sera aussi [98].

Le vieillissement est associé à des modifications corporelles importantes. Par exemple, la quantité de graisse sous-cutanée diminue avec l'âge alors que la quantité de graisse dans la cavité abdominale augmente accroissant alors le risque de maladie cardio-vasculaire ou de crise cardiaque [99]. L'augmentation de graisse abdominale est aussi liée à l'insulino-résistance et constitue donc un facteur de risque du diabète de type 2 [100]. Des changements concernant les muscles squelettiques sont également une conséquence des modifications survenant dans les métabolismes de la synthèse et de la dégradation des protéines dues au vieillissement [101]. De même, le vieillissement est également lié à un déclin des fonctions endocrines, comme la diminution de la sécrétion d'hormone de croissance [102, 103]. Ces diminutions peuvent également conduire à des pathologies caractéristiques du vieillissement.

6.2.2 Vieillissement cellulaire

De nombreux processus cellulaires sont altérés par le vieillissement.

Dans un premier temps, les cellules ont une capacité de prolifération limitée appelée sénescence cellulaire [104]. Une cellule sénescente perd sa capacité de croissance et la réplication de l'ADN. La sénescence cellulaire est associé au vieillissement puisque le nombre de cellules sénescentes augmente avec l'âge [105]. De plus, ces cellules montrent également une expression différentielle du patrimoine génétique notamment des gènes contrôlant le cycle cellulaire mais également de gènes codant pour des protéines qui peuvent modifier les caractéristiques et le métabolisme de la cellule [106-108]. L'impossibilité de se

répliquer pour les cellules sénescentes permet de mieux lutter contre le développement de cancer en évitant la prolifération de cellules endommagées et d'assurer l'homéostasie du tissu. Mais avec l'âge et la diminution des capacités de réparation des lésions, de telles cellules favorisent le vieillissement (Figure 31). Plusieurs revues recensent les modifications causées par la sénescence cellulaire [105, 108].

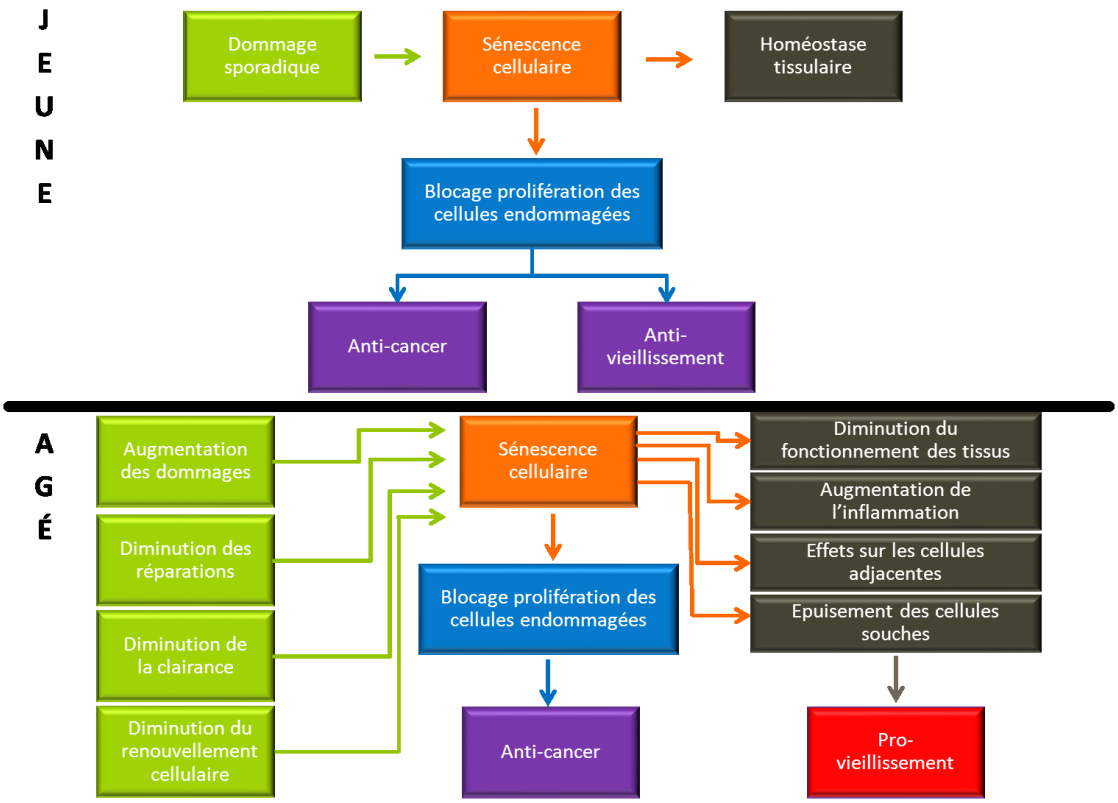


Figure 31. Comparaison des conséquences de la sénescence cellulaire chez les organismes jeunes et âgés. Chez les jeunes organismes, la sénescence cellulaire empêche la prolifération de cellules endommagées et constitue un mécanisme de prévention contre le cancer et le vieillissement qui garantit l'homéostasie des tissus. Au contraire, chez les organismes âgés, l'augmentation des dommages et la diminution du renouvellement cellulaire conduit à une accumulation de cellules sénescentes. Cette accumulation a des effets délétères sur l'homéostasie du tissu, favorisant le vieillissement. (Tirée de [109])

Le corps possède des réserves de cellules souches permettant la régénération des tissus. Les lésions auxquelles sont soumises les cellules entraînent la diminution de ces réserves et les tissus deviennent alors plus vulnérables aux dégâts. Une diminution trop rapide de ces stocks de cellules souches accélère le vieillissement.

De plus, les cellules possèdent des mécanismes permettant de s'assurer de la bonne qualité, de la stabilité voire de la destruction des protéines essentielles à leur fonctionnement. Or, le vieillissement conduit à des altérations de ces mécanismes pouvant conduire à des maladies [110]. En effet, différents stresses entraînent le mauvais repliement ou le dépliement des protéines. Les mécanismes dont le rôle est soit de détruire soit de replier convenablement les protéines défectueuses deviennent moins efficaces avec le temps. Cela conduit alors à une accumulation de protéines défectueuses ayant un effet toxique et favorisant le vieillissement.

Enfin, la cellule est en permanence agressée par des espèces réactives oxygénées (ROS, Reactive Oxygen Species en anglais). Bien que des systèmes de défense contre ces molécules existent, le vieillissement diminue l'efficacité de ces mécanismes. De plus, les ROS endommagent les mitochondries et engendrent des processus dégénératifs accélérant le vieillissement [111-113].

Toutes ces modifications apparentes au niveau cellulaire ont des explications moléculaires dues à des perturbations de l'ADN.

6.2.3 Perte de fonction moléculaire

Au cours du temps, le matériel génétique se détériore et de nombreuses lésions, comme des mutations, des translocations ou l'intégration de virus dans des séquences de gènes touchent l'ADN. Toutes ces altérations peuvent affecter des fonctions essentielles et empêcher le bon fonctionnement des cellules. Si ces cellules ne sont pas éliminées par apoptose ou par sénescence, elles peuvent compromettre l'intégrité du tissu ou son renouvellement. Afin de se prémunir contre ces changements, de nombreux mécanismes de réparation de l'ADN existent. Des défaillances dans ces systèmes de défense peuvent accélérer le vieillissement et favoriser l'apparition de pathologies ou le développement de cancers [114, 115]. De plus, des défauts dans la structure du noyau de la cellule peuvent favoriser cette instabilité génomique [116].

Une autre altération liée au vieillissement observable au niveau moléculaire affecte les télomères. Les télomères sont les régions terminales des chromosomes et subissent d'importantes modifications avec le temps. En effet, les enzymes dont la fonction est de répliquer l'ADN, les ADN polymérase, ne recopient pas entièrement les télomères. Même si des enzymes ont pour fonction la réplication de ces portions d'ADN, elles ne sont pas exprimées dans toutes les cellules. Par conséquent, la longueur des télomères diminue avec le temps. Cette diminution est à l'origine de la sénescence cellulaire. De plus, des anomalies

dans ces régions chromosomiques conduisent à des maladies caractérisées par une perte de la capacité de régénération des tissus comme la fibrose pulmonaire par exemple [117].

Enfin, de nombreuses altérations épigénétiques surviennent au cours du temps et peuvent favoriser l'instabilité génomique, influencer les mécanismes de méthylation de l'ADN et des, histones ou la conformation de la chromatine. Elles peuvent également modifier le niveau d'expression de certains gènes entraînant alors l'accélération du vieillissement [118, 119].

6.3 Vieillissement cutané

Comme tout autre organe du corps, la peau est sujette aux altérations engendrées par le temps. Le vieillissement cutané se traduit par des modifications observables à l'œil nu et le prélèvement d'échantillon de peau est plus facilement réalisable que pour d'autres tissus. Etant donné que ce tissu est considéré comme un bon modèle du vieillissement général et compte tenu de sa facilité d'étude, un intérêt particulier a été accordé au vieillissement cutané afin, d'une part, de minimiser les conséquences visibles des altérations de cet organe et, d'autre part, de comprendre les mécanismes responsable du vieillissement général.

6.3.1 Caractérisation du vieillissement cutané

Le vieillissement cutané, résultant d'un déséquilibre entre les lésions cellulaires et les mécanismes de défense, se traduit histologiquement par des changements majeurs dans la composition et l'organisation du tissu. Toutes les couches sont soumises à des modifications.

Dans l'épiderme, tous les types de cellules sont concernés. Au cours du temps, les cellules de la couche basale sont moins uniformes et leur taille augmente. Les kératinocytes changent de forme, ils deviennent plus courts et plus gras. Le nombre de mélanocytes fonctionnels et de cellules de Langherans diminue dans l'épiderme conduisant respectivement à une pigmentation non uniforme et une perte d'efficacité du système immunitaire. Ces dernières cellules ont moins de dendrites et donc une capacité moindre à piéger les agents pathogènes. De plus, la jonction dermo-épidermique est modifiée du fait de la réduction des zones de contact entre ces deux couches et donc des échanges de nutriments et de métabolites.

Le derme s'atrophie, le nombre de fibres de collagènes s'amoindrit et celles-ci deviennent plus épaisses et moins organisées. Le renouvellement de ces fibres est également moins important. Les fibres d'élastines sont également moins nombreuses, se renouvellent moins et

sont dégradées. Enfin, le nombre de fibroblastes diminue et la vascularisation du derme devient moins importante.

L'hypoderme subit aussi quelques modifications. Le volume de graisse dans cette couche de la peau diminue avec l'âge et entraîne des altérations de la fonction de thermorégulation. Les personnes âgées deviennent alors plus vulnérables aux hypothermies pour cette raison.

Tissu	**Modification observée**
Epiderme	Aplatissement de la jonction dermo-épidermique Diminution du nombre de mélanocytes actifs Diminution du nombrede cellules de Langherans Diminution de la capacité de reépithélisation Augmentation du nombre de pores
Derme	Atrophie Diminution de la vascularisation et du nombre et de la qualité des cellules Diminution du nombre de glandes sudoripares et modification de leur forme Diminution du nombre de vaisseaux et de terminaisons nerveuses
Hypoderme	Modification de la répartition des graisses sous-cutanées Diminution du volume général

Tableau 4. Caractéristiques du vieillissement cutané [120].

Tous ces changements au sein de la structure de l'organe se traduisent visiblement par l'apparition de rides, un relâchement du tissu, une sécheresse de la peau ou encore des troubles pigmentaires [121, 122].

Le processus multifactoriel du vieillissement cutané s'explique en partie par des facteurs environnementaux et comportementaux [122, 123].

6.3.2 Facteurs impactant le vieillissement cutané

Plusieurs facteurs environnementaux influent sur le vieillissement cutané. L'exposition au soleil, et en particulier aux rayons ultra-violets (UV), joue un rôle majeur dans le vieillissement de la peau. En effet, les rayons UVA sont absorbés par des molécules ayant un rôle photosensibilisateur et induisant une forte production d'espèces réactives de l'oxygène qui mettent à mal les défenses anti-oxydantes de la peau [124-126]. De plus, les rayons UVB pénètrent au niveau de l'épiderme et entraînent des dommages touchant l'ADN situé dans les noyaux des kératinocytes et des mélanocytes. Ces rayons entraînent également l'augmentation de l'expression des métalloprotéinases responsables de la dégradation des molécules de la matrice extracellulaire. L'épiderme devient irrégulier, atrophié ou hyperplasique. Le nombre de cellules de Langherans diminue alors que le nombre de mélanocytes hyperplasiques

augmente. Le derme subit également de nombreuses altérations, notamment au niveau de sa vascularisation. Les vaisseaux situés juste sous l'épiderme ont tendance à disparaître alors que les vaisseaux plus profonds se dilatent. Dans le même temps, le nombre de fibres de collagène diminue et le tissu élastique perd ses propriétés. Le vieillissement dû à l'exposition à ces rayonnements se manifeste par une peau plus épaisse, rugueuse, hyperlaxe et jaunâtre. Il entraîne également l'apparition de rides ou de taches pigmentaires, appelées lentigines, témoins de l'altération des mélanocytes (Figure 32). Les dommages engendrés par l'exposition aux rayons UV s'additionnent aux modifications histologiques de la peau et accélèrent par conséquent le vieillissement cutané.

Figure 32. Effet de l'exposition au soleil sur le vieillissement cutané chez deux jumelles. La jumelle de droite s'exposait en moyenne 10 heures de plus que la jumelle de gauche. (Tirée de [127])

La consommation de tabac est également un facteur accélérant le vieillissement cutané [128]. En effet, le tabac engendre une diminution de la circulation sanguine dans les capillaires du derme et par conséquent une baisse de l'apport de nutriment et d'oxygène. La nicotine a un rôle vasoconstricteur et son absorption favorise ainsi l'apparition de rides. De plus, le nombre de fibres de collagène et d'élastine diminue et la peau est alors moins élastique et moins ferme. Enfin, le tabac déforme les kératinocytes et rend la peau plus rugueuse. Les manifestations les plus visibles de la consommation de tabac sur le vieillissement sont l'apparition de rides (Figure 33). La formation de rides engendrée par le tabac est positivement corrélée à la quantité consommée [128].

Figure 33. Effet de la consommation de tabac sur le vieillissement cutané chez deux jumelles. La jumelle de gauche a fumé 20 ans de plus que la jumelle de droite. (Tirée de [127])

Enfin, d'autres facteurs influencent le vieillissement cutané. La consommation d'alcool est également connue comme étant un élément aggravant. En revanche, l'indice de masse corporelle est corrélé négativement avec le vieillissement [127, 129] : plus l'IMC est important, moins la peau semble vieille (Figure 34).

Figure 34. Effet de l'indice de masse corporelle sur le vieillissement cutané chez deux jumelles. La jumelle de gauche a un indice de masse corporelle supérieur de 14,7 points par rapport à celui de la jumelle de droite. (Tirée de [127])

Les œstrogènes sont également connus pour leur impact sur la peau [130-136]. Plusieurs études ont mis en évidence les changements observés sur la peau de femmes présentant des carences en œstrogènes ou ménopausées. En effet, la diminution du taux de ces hormones est liée, entre autres, à une perte en fibres de collagène. Par conséquent, la structure du tissu est moins bien assurée et l'apparition de rides est favorisée [137, 138]. Les traitements de

remplacement hormonaux post-ménopause sont connus pour réduire les conséquences de la baisse des œstrogènes et atténuer le vieillissement cutané [137, 139-141] (Figure 35).

Figure 35. Effet d'un traitement de substitution hormonal post ménopause chez deux jumelles. La jumelle de droite a reçu un traitement de remplacement 22 ans de plus que la jumelle de gauche. (Tirée de [127])

6.3.3 Génétique du vieillissement cutané

Les facteurs génétiques influençant le vieillissement cutané sont, à l'heure actuelle, encore largement méconnus. Les études tentant d'identifier ces facteurs sont donc très importantes, non seulement pour mieux appréhender les mécanismes du vieillissement cutané mais aussi à plus large échelle ceux du vieillissement général. En effet, la peau est considérée comme un modèle de vieillissement général, et des échantillons de peau sont même utilisés pour évaluer l'état de vieillissement du système nerveux central [142].

De nombreuses études d'expression des gènes ont été menées (parmi lesquelles [143-146]). Ainsi, en 2006, Lener *et al.* établissent une liste de 105 gènes dont le taux d'expression varie d'un facteur supérieurs à 1,7 entre des échantillons de peau prélevés chez de jeunes garçons de 3-4 ans ou des hommes de 68 à 72 ans [147]. Les gènes ainsi mis en évidence, régulés positivement ou négativement par le processus de vieillissement, sont impliqués dans de nombreux processus cellulaires, et en particulier pour près de 20% d'entre eux, dans le métabolisme, la liaison à l'ADN et l'activation ou la répression de sa transcription. Une étude menée cette fois sur 10 jeunes femmes de 19 à 20 ans et 10 femmes plus âgées (63 à 67 ans) a permis d'obtenir des résultats assez similaires[148]. En effet, ce travail a mis en évidence la régulation négative de gènes impliqués dans les processus biologiques de synthèse des lipides

ou dans la différentiation des cellules de l'épiderme ainsi que la régulation positive des gènes contrôlant les processus de réponse à l'inflammation et des gènes de l'élastine. Même si ces 2 études ont donné des résultats similaires, il est intéressant de savoir si les mécanismes de vieillissement cutané sont comparables chez l'homme et la femme. Dans une publication de 2012, Makrantonaki et al. [149] montrent qu'il existe une différence de variation du taux d'expression de certains gènes lors de l'étude du vieillissement cutané entre les hommes et les femmes. En effet, chez la femme et chez l'homme, l'expression de respectivement 523 et 401 gènes étaient significativement régulée, parmi lesquels seulement 39 étaient communs à l'homme et la femme. Cette étude a aussi révélé une possible implication de la voie de signalisation Wnt puisque 4 gènes appartenant à cette voie de signalisation sont régulés négativement chez les personnes âgées. Ce résultat est particulièrement intéressant puisque la voie de signalisation Wnt est suspectée d'agir dans le processus de vieillissement, même si son rôle reste non élucidé et même controversé [150, 151].

Au-delà de ces études d'expression génique, des études gènes-candidats ont aussi été mises en œuvre pour identifier les gènes impliqués dans le vieillissement cutané. Le gène *MC1R* (MelanoCortin-1 Receptor) a notamment été ciblé par ses études [152, 153]. L'objectif de la première étude [152] était d'étudier la possible association entre la sévérité du photo-vieillissement cutané facial et des allèles du gène *MC1R*. En effet, les associations de ce gène avec des phénotypes considérés à risque tels que les cheveux roux, la peau claire et les tâches de rousseurs [154-159], avec les lentigines solaires [158, 160] ou encore avec le risque de développer des mélanomes [161-165] ont notamment été mises en évidence. Cette étude a démontré que la possession de 2 allèles du gène *MC1R* augmente le risque de présenter un photo-vieillissement cutané facial sévère, et ce risque est encore augmenté lorsqu'il s'agit de 2 allèles majeurs. La seconde étude s'intéresse quant à elle au vieillissement cutané péri-orbital [153] en étudiant sa prédictibilité. Là encore, les allèles du gène *MC1R* apparaissent comme étant de bons prédicteurs du photo-vieillissement cutané.

Plus récemment, des études d'association « génome entier » ont été réalisées. La première date de 2012 et a été réalisée dans notre laboratoire [166]. L'objectif de cette GWAS était de tenter de déterminer les facteurs génétiques impactant la sévérité du vieillissement cutané. Le SNP rs322458 a été identifié comme significativement associé avec le score global de photo-vieillissement (déterminé par des dermatologues), mais aussi avec les scores évaluant les rides et l'affaissement de la paupière. Ce SNP était en déséquilibre de liaison avec des SNPs introniques du gène *STXBP5L* exprimé dans la peau et avec des SNPs augmentant

l'expression du gène *FBXO40* dans la peau. Ces 2 gènes n'avaient jamais été mis en évidence dans les mécanismes contrôlant le vieillissement cutané, cependant quelques pistes pourraient expliquer leur implication. Ainsi, l'implication du gène *STXBP5L*, identifié comme étant associé au cancer, s'explique rationnellement puisque les cancers de la peau sont intimement liés au vieillissement cutané. Le gène *FBXO40* pourrait quant à lui jouer un rôle dans les mécanismes contrôlant la formation des rides et l'affaissement de la paupière puisque ce gène est lié à la voie de signalisation IGF1 (Insuline-like Growth Factor 1), connue pour intervenir dans l'inflammation et la myogenèse. Une deuxième GWAS a été publiée très récemment [167], se consacrant cette fois uniquement à l'affaissement de la paupière. Son objectif était de tenter de déterminer les facteurs intrinsèques et extrinsèques contrôlant ce phénomène, fréquemment observé dans le vieillissement cutané. Ce travail a identifié l'allèle C du SNP rs11876749 comme étant protecteur du risque d'affaissement de la paupière. Ce SNP est notamment localisé à proximité du gène *TGIF1* (TGFB-Induced Factor homeobox1), capable d'induire la voie du TGF- β (Tumor Growth Factor β) déjà connu pour son implication dans le vieillissement cutané [168]. Enfin, une troisième étude génome entier s'est intéressée à l'identification de possibles gènes protégeant le visage du vieillissement cutané [169]. Les résultats de cette GWAS ont permis d'identifier 3 possibles gènes promoteurs de jeunesse cutanée. Le premier, *KCND2*, est exprimé dans la peau mais est sans relation évidente avec le vieillissement cutané. Le second, *DIAPH1*, est associé à l'insuffisance ovarienne prématurée [170, 171] et donc possiblement au vieillissement cutané puisque le lien entre celui-ci et la ménopause a déjà été démontré. Enfin, le troisième, *EDEM1*, peut être le plus intéressant puisqu'il a été mis en évidence comme contrôlant la durée de vie de certains insectes [172], mais, cependant, pas de l'Homme.

7 Objectifs de ma thèse

J'ai intégré l'équipe du Professeur Jean-François Zagury au sein du laboratoire Génomique Bioinformatique et Applications (GBA) en tant que stagiaire en avril 2010 afin de me familiariser avec l'exploitation bioinformatique et statistique des données génomiques. A l'heure où les études d'association génome entier sont devenues un outil incontournable dans la recherche de facteurs génétiques impliqués dans certaines pathologies et profitant du fort savoir-faire de l'équipe dans ce domaine, j'ai décidé de m'engager dans une thèse afin d'acquérir et d'appliquer les compétences nécessaires à ce genre d'étude.

Au cours du temps, mon projet de thèse s'est dessiné autour de deux axes majeurs :

1. Exploiter le savoir-faire acquis dans le cadre d'études d'association génome entier se focalisant sur différents marqueurs du vieillissement cutané ;

2. Acquérir les méthodes d'analyse des voies de signalisation biologiques et les mettre en œuvre sur ces mêmes données afin de compléter les résultats obtenus lors des GWAS.

Deuxième partie
Matériels et Méthodes

1 Etudes d'association génome-entier

1.1 Population étudiée

1.1.1 La cohorte SU.VI.MAX

L'étude SU.VI.MAX [173, 174] (SUpplémentation en VItamines et en Minéraux AntioXydants) est une étude longitudinale conduite en France sur une population adulte d'âge moyen. L'étude SU.VI.MAX avait été initialement développée pour évaluer l'effet d'une supplémentation nutritionnelle quotidienne sur la réduction des problèmes de santé publique, tels que les cancers ou maladies cardio-vasculaires dans les pays industrialisés. La cohorte inclut 13 017 volontaires d'âge moyen avec un spectre représentatif de situations sociodémographiques [174]. Le protocole de l'étude SU.VI.MAX a été approuvé par le Comité d'éthique de l'Hôpital Paris-Cochin (CCPPRB n° 706) ainsi que par le "Comité National Informatique et Liberté" (CNIL n°334641). L'étude a été menée conformément aux principes de la Déclaration d'Helsinki.

1.1.2 Cohorte des femmes étudiées pour le vieillissement cutané

De l'automne 2002 et jusqu'à l'hiver 2003, les femmes de la cohorte SU.VI.MAX vivant en région parisienne ont été interrogées sur leur volonté de participer à une étude sur le vieillissement cutané. Parmi elles (n = 2 257), 570 femmes dont l'âge était compris entre 44 et 70 ans ont fourni leur consentement éclairé pour faire partie de ce projet. Il leur a alors été demandé de suivre des consignes relatives aux soins cutanés. En particulier, les patientes ne devaient appliquer aucun produit (cosmétique ou détergent) sur leur visage pendant les douze heures précédant la visite programmée pour récolter les informations nécessaires à l'étude. Le jour de cette visite, les patientes ont dans un premier temps dû répondre à un questionnaire relatif à leurs habitudes d'exposition au soleil. Ensuite, trois photographies à haute résolution (2008 x 3032 pixels) de leur visage (une de face et une pour chaque profil) ont été prises à l'aide d'un appareil numérique Kodak DCS 760 équipé d'un objectif de 105 millimètres. Ces photographies standardisées ont été réalisées en suivant un protocole précis. D'une part, l'appareil photographique était placé sur un monopode et les femmes étaient assises sur une chaise spécialement conçue pour l'étude afin d'assurer une position identique du visage de

chacune des patientes par rapport à l'objectif de l'appareil. D'autre part, deux lampes reproduisant le spectre continu de la lumière du jour ont été placées de façon symétrique à 45° de chaque profil du visage afin de permettre un éclairage standardisé pour chaque photographie. Enfin, un échantillon sanguin a été prélevé pour chacune des 570 femmes.

1.2 Phénotypes analysés

Pour chacune des femmes, un dermatologue a analysé les photographies afin d'évaluer la sévérité du photo-vieillissement cutané et de plusieurs autres indicateurs cliniques du vieillissement cutané sur le visage.

Le photo-vieillissement a été estimé à l'aide d'une échelle ordinale composée de six grades développée par C. Larnier [175]. Chacun des six grades de cette échelle est représenté par trois photographies de référence illustrant la diversité et la variété des troubles pigmentaires, des rides et du relâchement de la peau (Figure 36).

Figure 36. Echelle d'évaluation du photo-vieillissement. (Adaptée de [175])

Outre le photo-vieillissement, le dermatologue a également évalué douze indicateurs cliniques du vieillissement cutané à l'aide d'échelles photographiques préalablement validées [176] :

Troubles pigmentaires	Lentigines sur le front	Grade de 0 à 4
	Lentigines sur les joues	Grade de 0 à 4
Rides	Rides inter-sourcilières	Grade de 0 à 5
	Rides de la patte d'oie	Grade de 0 à 5
	Rides sous les yeux	Grade de 1 à 5
	Rides fines sur les joues	Grade de 0 à 2
	Rides d'expression des joues	Peu/très marquées
	Rides du contour de la bouche	Grade de 1 à 4
Relâchement	Poches sous les yeux	Absence/présence
	Relâchement de l'ovale du visage	Grade de 0 à 5
	Relâchement des paupières supérieures	Grades de 0 à 5
	Sillons naso-géniens	Grade de 0 à 5

Tableau 5. Ensemble des douze indicateurs cliniques du vieillissement cutané évalués.

A partir de ces douze évaluations, trois scores globaux mesurant la sévérité des troubles pigmentaires (Tableau 6), des rides (Tableau 7) et du relâchement (Tableau 8) ont été déterminées à l'aide d'une ACP et d'une régression linéaire.

Score global de lentigines		Coefficient
Lentigines sur le front	si grade = 0	+ 0
	si grade = 1	+ 1,25
	si grade = 2	+ 2,5
	si grade = 3	+ 3,75
	si grade = 4	+ 5
Lentigines sur les joues	si grade = 0	+ 0
	si grade = 1	+ 1,25
	si grade = 2	+ 2,5
	si grade = 3	+ 3,75
	si grade = 4	+ 5
Score		**= Somme**

Tableau 6. Calcul du score global de lentigines.

Chacun des scores globaux, des douze grades et le grade de photo-vieillissement ont été analysés.

Score global de rides		Coefficient
Constante		-0,64
Rides inter-sourcilières	si grade < 2	+ 0
	si grade = 2	+ 0,44
	si grade = 3	+ 0,88
	si grade = 4	+ 1,22
	si grade = 5	+ 1,66
Rides de la patte d'oie	si grade < 2	+ 0
	si grade = 2	+ 0,54
	si grade = 3	+ 1,08
	si grade = 4	+ 1,62
	si grade = 5	+ 2,16
Rides sous les yeux	si grade < 3	+ 0
	si grade = 3	+ 0,64
	si grade = 4	+ 1,28
	si grade = 5	+ 1,92
Rides fines sous les joues	si grade =0	+ 0
	si grade = 1	+ 0,70
	si grade = 2	+ 1,4
Rides d'expression des joues	si très marquées	+ 1,06
Rides du contour de la bouche	si grade =0	+ 0
	si grade = 1	+ 0,42
	si grade = 2	+0,84
	si grade = 3	+ 1,26
	si grade = 4	+ 1,68
Score		**= Somme**

Tableau 7. Calcul du score global de rides.

Score global de relâchement		Coefficient
Poches sous les yeux	si présence	+ 0,87
Relâchement de l'ovale du visage	si grade < 3	+ 0,93
	si grade = 3,	+ 1,86
	si grade > 3	+ 2,79
Affaissement de la paupière supérieure	si grade < 3	+ 1,07
	si grade = 3	+ 2,14
	si grade > 3	+ 3,21
Sillons naso-géniens	si grade < 3	+ 0,78
	si grade = 3	+ 1,56
	si grade = 4	+ 2,34
	si grade > 4	+ 3,12
Score		**= Somme**

Tableau 8. Calcul du score global de relâchement.

1.3 Covariables utilisées dans l'analyse statistique

Plusieurs caractéristiques susceptibles d'avoir une influence sur le vieillissement cutané ont été prises en compte dans les analyses statistiques :

- L'âge en années
- Le statut tabagique en trois classes (jamais fumeur, ancien fumeur et actuellement fumeur)
- L'IMC catégorisé en trois classes selon les recommandations de l'Organisation Mondiale de la Santé : en sous poids ou normal si l'IMC est strictement inférieur à 25 kgm^{-2}, en surpoids si l'IMC est compris entre 25 kgm^{-2} et 30 kgm^{-2}, et obèse si l'IMC est supérieur à 30 kgm^{-2}
- Le statut hormonal en trois classes : non-ménopausée, ménopausée et suivant une thérapie hormonale de substitution et ménopausée sans traitement hormonal de substitution
- L'intensité de l'exposition au soleil au cours de la vie évaluée à partir d'un auto-questionnaire. Ce score est une combinaison de cinq items pondérés suivant leur contribution respective à cette évaluation : exposition volontaire au soleil, exposition du corps et du visage, exposition durant les heures les plus chaudes de la journée, auto-évaluation de l'intensité de l'exposition au soleil au cours de la vie et importance accordée au bronzage. La conception, la validation et la description de ce score ont déjà été renseignées auparavant [177].

1.4 Génotypage

Sur les 570 femmes ayant accepté de participer à cette étude, 41 femmes n'ont pas été génotypées. En effet, 10 patientes ont été exclues puisque d'origine non caucasienne, 18 autres en raison d'opérations de chirurgie esthétique, 13 femmes avaient un échantillon d'ADN trop endommagé ou pas suffisamment concentré. Les 529 femmes restantes ont été génotypées en utilisant des puces de génotypage Illumina Infinium HumanOmni1-Quad renseignant 1 140 419 polymorphismes. L'ADN génomique (250 ng) de chacune des femmes a alors été amplifié, fragmenté, dénaturé et hybridé sur une puce pendant au moins 16 heures à 48°C. Les fragments hybridés de manière non spécifique ont été éliminés après lavage. Les fragments restants, hybridés de façon spécifique, ont été marqués par fluorescence et détectés

en utilisant un scanner Illumina IScan. Les intensités normalisées ont ensuite été interprétées en génotypes en utilisant la version 1.6.3 du logiciel GenomeStudio d'Illumina. Pour la suite des analyses, nous nous sommes focalisés uniquement sur les SNPs. Par conséquent, 91 706 CNVs génotypés ont été exclus ainsi que 2 182 SNPs situés sur le chromosome Y dans la mesure où la population étudiée est uniquement composée de femmes.

1.5 Contrôle qualité du génotypage

Les données brutes de génotypage ont été analysées à l'aide de la version 1.6.3 du logiciel GenomeStudio développé par Illumina. Tout d'abord, 9 individus ont été éliminés de l'étude en raison d'un taux de SNPs génotypés par individu (call rate en anglais) inférieur à 95%. Ensuite, les SNPs avec une proportion d'individu génotypés (call frequency en anglais) inférieur à 99% ont été à nouveau inférés. Suite à cela, les individus dont le call rate était inférieur à 98% ont été exclus des analyses. Ce protocole correspond aux recommandations faîtes par Illumina afin de minimiser les erreurs d'inférence et permettant de les corriger manuellement si nécessaire (http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf). Ensuite, les 56 479 SNPs présentant un taux de données manquantes supérieur à 2% ont été exclus. Afin de limiter les erreurs de génotypage, les 191 123 SNPs dont la fréquence de l'allèle mineur (MAF, Minor Allele Frequency en anglais) était inférieure à 1% n'ont pas été retenus pour l'analyse. Enfin, un écart trop important à l'équilibre d'Hardy-Weinberg pouvant être la conséquence de problèmes de génotypage, 3 866 SNPs dont la p-valeur du test exact d'adéquation à l'équilibre d'Hardy-Weinberg était inférieure à 5.0×10^{-3} ont également été exclus. Finalement, 795 063 SNPs ont été analysés.

1.6 Etude de la stratification de la population

Afin de corriger une possible stratification au sein de la cohorte, les génotypes ont été analysés par l'intermédiaire d'une analyse en composantes principales implémentée dans le logiciel EIGENSTRAT issu de la suite EIGENSOFT [72]. Une première passe, réalisée en incluant deux populations asiatiques de HapMap (Chinois de Pékin, CHB, et Japonais de Tokyo, JPT) et deux populations africaines (Yoruba du Nigéria, YRI, et Masaï du Kenya, MKK), ont permis d'identifier cinq individus atypiques. Une deuxième passe, réalisée sur la cohorte en excluant les individus éliminés lors de la première passe, a conduit à l'identification de treize individus atypiques supplémentaires. Au total, 18 individus ont été exclus de l'analyse lors de la stratification. Finalement, une dernière passe effectuée avec les

individus restant a été réalisée afin de déterminer les deux premiers vecteurs propres de la stratification inclus dans les analyses statistiques afin de tenir compte de la structure de la population.

1.7 Analyses statistiques

Sur les 570 femmes inclues initialement dans l'étude, 68 d'entre elles ont été exclues : 10 étaient visiblement d'origine non caucasienne, 13 présentaient un échantillon d'ADN endommagé ou trop peu concentré, 18 avaient subi des opérations de chirurgie esthétique, 9 ont été éliminées à l'issue du contrôle qualité et 18 ont été identifiées comme étant des individus atypiques lors de l'étude de la stratification. Ainsi, 502 femmes ont été inclues dans les analyses.

Dans un premier temps, la population a été décrite en fonction de la sévérité de chaque grade en utilisant des analyses de variance (ANOVA) pour les variables quantitatives et des tests du χ^2 pour les variables qualitatives. Ensuite, pour chacun des 795 063 SNPs, les associations entre les génotypes et chacun des grades et scores globaux ont été mesurées en utilisant des régressions logistiques multivariées pour les phénotypes dichotomiques et des régressions linéaires multivariées pour les phénotypes quantitatifs à l'aide du logiciel PLINK [178]. Dans chacun des cas, les différentes covariables ainsi que les deux premiers vecteurs propres de la stratification ont été inclus dans le modèle.

La correction des tests multiples a été prise en compte pour chaque phénotype analysé. Les p-valeurs d'association ont été ajustées selon la méthode de Bonferroni et comparées au seuil de significativité génome-entier ($5{,}0 \times 10^{-8}$). De plus, le FDR a été mesuré pour chacun des SNPs par l'intermédiaire du calcul de la q-valeur. Chaque SNP ayant une q-valeur inférieure à 25% était considéré comme significatif.

Les données imputées ont été analysées suivant la même méthodologie à l'aide de la version 2 du logiciel SNPTEST [179, 180].

1.8 Déséquilibre de liaison

Pour chaque SNP identifié lors des études d'association, tous les SNPs en fort déséquilibre de liaison avec ce polymorphisme ($r^2 > 0{,}8$) dans la population européenne du projet 1000 Genomes (EUR, 1000 Genomes phase 1 integrated release version 3,

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/) ont été identifiés. Seule la liste de SNPs distincts a été conservée.

1.9 Imputation

Tout d'abord, les données ont été phasées à l'aide du logiciel SHAPEIT2 [62]. Les données phasées ont ensuite été imputées en utilisant le logiciel IMPUTE2 [68]. Les génotypes de la population européenne du projet 1000 Genomes (EUR, 1000 Genomes phase 1 integrated release version 3, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/) ont été utilisées comme panel de référence. Afin de réduire l'incertitude, les SNPs dont la MAF est inférieure à 1% dans cet échantillon n'ont pas été éliminés du panel de référence. Lorsque cela était pertinent, les allèles du HLA a été réalisée grâce au logiciel SNP2HLA [181].

1.10 Exploration bioinformatique

Pour chaque SNP d'intérêt, nous avons exploré différentes bases de données afin d'établir d'éventuelles associations avec un mécanisme biologique. Plusieurs bases de données d'expression, disponibles publiquement, permettent d'établir des corrélations entre polymorphisme et variation du niveau d'expression des ARNm : Genevar (http://www.sanger.ac.uk/resources/software/genevar/), mRNA by SNP Browser (http://www.sph.umich.edu/csg/liang/asthma/), et GHS-Express (http://genecanvas.ecgene.net/uploads/ForReview/).

Des informations sont également disponibles sur des données de méthylation (Genevar, http://www.sanger.ac.uk/resources/software/genevar/) ; de polyadénylation (PolyApred, http://www.imtech.res.in/raghava/polyapred/submission.html) ; de sites de fixation de facteurs de transcription (regulomedb, http://regulomedb.org/ dérivée de la base de données TRANSFAC database) ; gènes ou cibles de miRNA (miRBAse, http://www.mirbase.org/; miRTarBase, http://mirtarbase.mbc.nctu.edu.tw; et MicroCosm Targets, http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/) ; et les données d'ENCODE (regulomedb, http://regulomedb.org/).

2 Analyse des voies de signalisation

Les recherches de voies de signalisation associées à un phénotype s'inscrivent dans la continuité des études d'association génome-entier. Ainsi, dans la cadre de notre nous avons entrepris une étude visant à repérer les voies de signalisation associés aux différents scores globaux de vieillissement cutané sur la face et au grade de photo-vieillissement. Pour ce faire, nous avons utilisé les résultats d'association des SNPs obtenus lors des GWAS.

2.1 Voies de signalisation analysées

Plusieurs bases de données ont pour vocation de compiler des voies de signalisations. Dans notre étude, nous avons utilisé une liste composée de 212 voies de signalisation élaborée à partir de la base de données KEGG [182] disponible avec la suite gratuite de logiciels GenGen (http://www.openbioinformatics.org/gengen/index.html). Afin de se prémunir contre le biais potentiel induit par la variabilité importante du nombre de gènes entre les différentes voies de signalisation, seuls 139 les pathways constitués d'au moins 20 gènes et d'au plus 200 gènes ont été analysés.

2.2 Assignation des SNPs aux gènes

Les 795 063 inclus dans l'analyse ont été assignés aux gènes selon leur position physique. Un SNP était considéré comme appartenant à un gène s'il était situé dans le gène ou au plus 5kb en amont ou en aval du gène. Lorsqu'un SNP respectait cette condition pour plusieurs gènes, ce polymorphisme était considéré comme situé dans chacun de ces gènes.

2.3 Assignation des p-valeurs aux gènes

Pour chacun des gènes, la p-valeur assignée à ce gène était la p-valeur individuelle la plus faible parmi celles des SNPs situés dans ce gène.

2.4 Description de la méthode GSEA / Analyse statistique

L'association des voies métaboliques avec le phénotype a été mesurée à l'aide de l'algorithme Gene Set Enrichment Analysis [183] implémenté dans la suite de logiciels GenGen (http://www.openbioinformatics.org/gengen/index.html).

Soit N le nombre de gènes inclus dans l'analyse. L'ensemble des p-valeurs r_j de chacun des gènes G_j est ordonnée de façon décroissante. Cette liste de p-valeurs ordonnées est notée $r_{(1)},\ldots,r_{(N)}$. Un score d'enrichissement *ES(S)* est calculé pour chaque voie métabolique S composée de N_H gènes :

$$ES(S) = \max_{1 \le j \le N} \left\{ \sum_{\substack{G_{j^*} \in S \\ j^* \le j}} \frac{\left|r_{(j^*)}\right|^p}{N_R} - \sum_{\substack{G_{j^*} \notin S \\ j^* \le j}} \frac{1}{N - N_H} \right\}$$

avec $N_R = \sum_{G_{j^*} \in S} \left|r_{(j^*)}\right|$ et p un paramètre permettant d'accorder un poids plus important aux faibles p-valeurs.

Ce score d'enrichissement tenant compte de la p-valeur la plus faible dans chaque gène, les gènes comportant un grand nombre de SNPs ont une probabilité plus importante de se voir assigner une p-valeur faible. De même, les voies métaboliques composées d'un grand nombre de gènes ont une plus grande chance d'inclure des gènes ayant des p-valeurs faibles. Afin de tenir compte de ces différences de tailles de gènes, le score d'enrichissement est normalisé à l'aide de permutations soit des phénotypes soit des SNPs. Pour chaque permutation notée π, le score d'enrichissement ES(S, π) est calculé comme précédemment pour chaque voie métabolique. Un score d'enrichissement peut alors être calculé pour chaque pathway S :

$$NES(S) = \frac{ES(S) - mean[ES(S,\pi)]}{SD[ES(S,\pi)]}$$

Où mean$[ES(S,\pi)]$ et SD$[ES(S,\pi)]$ désignent respectivement la moyenne et l'écart type de la distribution des scores d'enrichissement obtenus en permutant.

La significativité des scores d'enrichissement et la correction des tests multiples sont effectués à l'aide des permutations. Une p-valeur empirique P_S est determinée pour chaque pathway S :

$$P_S = \%\ permutations\ telles\ que\ NES(S,\pi)\ \geq NES(S)$$

La procédure utilisée correction des tests multiples est le FDR. En notant NES* le score d'enrichissement obtenu avec les données non permutées pour une voie métabolique, le FDR est donné par :

$$FDR = \frac{\%\ de\ couples\ (S,\pi)\ tels\ que\ NES(S,\pi) \geq NES^*}{\%\ de\ voies\ métaboliques\ S\ telles\ que\ NES(S)\ \geq NES^*}$$

2.5 Analyse statistique

Afin de rechercher les voies métaboliques associées au grade de photo-vieillissement et aux différents scores mesurant la sévérité des troubles pigmentaires, des rides et du relâchement, les p-valeurs d'association en mode génotypique des 795 063 SNPs avec chacun de ces phénotypes obtenues lors des études génome-entier sur une cohorte de 502 femmes caucasiennes. Ces p-valeurs ont été obtenues à l'aide de régressions linéaires multivariées en incluant l'âge, le statut tabagique, le statut hormonal, l'exposition au soleil, l'IMC et les deux premiers vecteurs propres déterminés lors de l'étude de la stratification comme covariables. Pour chacune des 139 voies de signalisation issues de KEGG analysées, un score d'enrichissement a été calculé à l'aide de l'algorithme GSEA. Afin de déterminer la significativité statistique de l'association des voies de signalisationavec les différents phénotypes et de corriger les tests multiples, un cycle de 1000 permutations sur les SNPs a été effectué. Les voies de signalisation dont le FDR était inférieur à 5% ont été considérées comme significativement associées au phénotype et celles dont le FDR était inférieur à 25% ont été considérées comme hypothétiquement associées au phénotype.

Troisième partie

Résultats

1 Etude “génome entier” sur une cohorte de femmes caucasiennes à la recherche de gènes associés à l’apparition des lentigines

A Genome-Wide Association Study in Caucasian Women suggests the involvement of HLA genes in the severity of lentigines on the face

Vincent Laville*, Sigrid Le Clerc*, Khaled Ezzedine, Randa Jdid, Lieng Taing, Toufik Labib, Cedric Coulonges, Damien Ulveling, Wassila Carpentier, Pilar Galan, Serge Hercberg, Frederique Morizot, Julie Latreille, Denis Malvy, Erwin Tschachler, Christiane Guinot, Jean-François Zagury

J. Invest. Dermatol **(en révision)**

Afin d’élucider les facteurs génétiques impliqués dans la sévérité des lentigines, nous avons entrepris une analyse « génome entier » sur 520 des 2257 femmes de la cohorte SU.VI.MAX [173]. Pour cela, nous avons étudié si la sévérité des lentigines, évaluée à l’aide de 3 scores (un grade de lentigines sur le front, un grade de lentigines sur les joues et un score global de lentigines sur le visage) [152, 184] était associée ou non à des gènes spécifiques.

Nous nous sommes tout d’abord intéressés aux associations génétiques pouvant être corrélées au grade de sévérité des lentigines sur le front. Nos résultats nous ont permis de mettre en évidence 2 SNPs en déséquilibre de liaison total dont la p-valeur d’association dans le mode génotypique ($P_{GENO} = 1,37 \times 10^{-8}$) était inférieure au seuil de significativité génome-entier (5×10^{-8}). Ces SNPs étaient localisés sur le chromosome 6, dans une partie intergénique de la région 6p22, les gènes les plus proches étant le pseudogène *RPL29P17* et le gène *MBOAT1*. Nous avons identifié 18 SNPs, tous situés dans la même partie intergénique de la région 6p22, en déséquilibre de liaison important ($r^2 > 0,8$) avec au moins un des deux SNPS repérés précédemment.

Aucune association significative n'a pu être mise en évidence entre des SNPs et le grade de sévérité des lentigines sur les joues.

Finalement, nous avons identifié plusieurs associations entre le score global des lentigines sur le visage et 9 SNPs (FDR < 25%). Ces 9 SNPs étaient localisés sur le chromosome 6 et répartis en 2 groupes indépendants, séparés de 11 Mb. Le premier groupe était formé des 2 SNPs précédemment identifiés comme étant significativement associés au grade de lentigines sur le front. Le second groupe, constitué de 7 SNPs situés dans la région HLA, était non seulement associé avec une modification de l'expression de plusieurs gènes de la région HLA, et en particulier *HLA-C*, mais aussi en fort déséquilibre de liaison avec l'allèle *HLA-C*0701*.

Cette étude nous a donc permis d'identifier des gènes significativement associés avec la sévérité des lentigines sur le visage. L'un des résultats majeurs de cette étude impliquant la région HLA est la confirmation du lien entre le système immunitaire et l'apparition de lentigines. En particulier, le gène *HLA-C*, acteur essentiel de l'immunité acquise et adaptative, semble jouer un rôle dans la sévérité des lentigines.

A Genome-Wide Association Study in Caucasian Women reveals the involvement of HLA genes in the severity of facial solar lentigines

Vincent Laville[1]*, Sigrid Le Clerc[1]*, Khaled Ezzedine[2,3], Randa Jdid[4], Lieng Taing[1], Toufik Labib[1], Cedric Coulonges[1], Damien Ulveling[1], Wassila Carpentier[5], Pilar Galan[2], Serge Hercberg[2,6], Frederique Morizot[4], Julie Latreille[4], Denis Malvy[2,7], Erwin Tschachler[8,9], Christiane Guinot[8,10], Jean-François Zagury[1]*[#].

[1] Équipe Génomique, Bioinformatique et Applications, Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, Paris, France;
[2] UMR U557, INSERM/U1125 INRA/CNAM, University Paris 13/Centre de Recherche en Nutrition Humaine Ile-de-France, Bobigny, France;
[3] Department of Dermatology, Hôpital Saint-André, Bordeaux, France;
[4] Chanel R&T, Department of Skin Knowledge & Women Beauty, Pantin, France;
[5] Plate-forme Post-Génomique P3S, Hôpital Pitié-Salpêtrière, Paris, France;
[6] Department of Public Health, Hôpital Avicenne, Bobigny, France;
[7] Department of Internal Medicine and Tropical Diseases, Hôpital Saint-André, Bordeaux, France;
[8] CE.R.I.E.S.[§], Neuilly sur Seine, France;
[9] Department of Dermatology, University of Vienna Medical School, Vienna, Austria;
[10] Computer Science Laboratory, University François Rabelais of Tours, Tours, France;

[§] CE.R.I.E.S. is a research centre on human skin founded by CHANEL.
*: These authors share an equal contribution to the work.

[#]: Correspondance to:

Jean-François Zagury,
292 rue Saint Martin, 75003 Paris, France,
Tel: 33 1 58 80 88 20, e-mail: zagury@cnam.fr

Short Tittle: *HLA* genes in the severity of solar lentigines

Abbreviations: GWAS, Genome Wide Association Study; eQTL, expression Quantitative Trait Loci; FDR, False Discovery Rate; HLA, Human Leukocyte Antigen; LD, Linkage Disequilibrium; MAF, Minor Allele Frequency; SL: Solar Lentigine; SNP, Single Nucleotide Polymorphism.

Abstract

We have performed a genome-wide analysis to identify genes potentially associated with solar lentigines in 502 middle-aged French women. In this study, we scored the severity of solar lentigines on the forehead, on the cheeks, and globally on the face. The associations between SNPs and each outcome variable were measured using linear regressions adjusted for potential confounding factors. Imputation of the 1000 genomes SNP/indels as well as HLA class I and II alleles was also performed.

Nine SNPs, gathered into two independent blocks of chromosome 6, with a False Discovery Rate below 25% when looking for associations with the facial solar lentigine score. The first block, in the 6p22 region, corresponded to intergenic SNPs and also exhibited an association with forehead solar lentigines ($p = 1.37 \times 10^{-8}$). The second block, within the 6p21 HLA region, was associated with a decreased *HLA-C* expression according to several eQTL databases. Interestingly, these SNPs were also in high linkage disequilibrium with the *HLA-C*0701* allele ($r^2 = 0.95$). An additional study on 38 candidate genes also revealed several signals in the *MITF* gene.

Overall, our results point to several mechanisms involved in the severity of facial solar lentigines including HLA/immunity and the melanogenesis pathway.

Introduction

Skin aging, a natural process taking place over time, is the result of both environmental (extrinsic) and chronological (intrinsic) factors. However, whereas chronological aging is a genetically determined process resulting in the natural degeneration of cell function and the loss of extracellular matrix with passage of time (Yaar and Gilchrest, 2007), extrinsic aging is mainly driven by chronic sun exposure resulting in the so called photoaging. The clinical phenotype of photoaging includes wrinkling, coarseness and lentigines (Bastiaens *et al.*, 2004; Ezzedine *et al.*, 2013). Recently, two genome-wide association studies (GWAS) have been performed highlighting the possible role of genes such as *DIAPH2* on skin youthfulness (Chang *et al.*, 2014) and of *STXBP5L* on global photoaging (Le Clerc *et al.*, 2013). Interestingly, in this latter study, the *STXBP5L* signal was found to be associated with both wrinkles and sagging but not with solar lentigines (SL). Indeed, SL are dark brown spots occurring on sun-exposed areas such as the face or the back of the hands and are considered as a good indicator of extrinsic skin aging. Several studies have described the histopathological features of SL (Andersen *et al.*, 1997) although to date, little is known about the contribution of genetics to SL. In particular, no genome-wide association study (GWAS) targeting SL has previously been reported.

In this context, we re-analyzed data from a previously published GWAS (Le Clerc *et al.*, 2013) to determine whether the severity of SL, using two grades and one validated score (Elfakir *et al.*, 2010; Ezzedine *et al.*, 2013), is associated or not with specific genes.

Results

Genotype data were obtained with the Illumina HumanOmni1-Quad BeadChips on a sample of 502 French middle-aged women from the SU.VI.MAX cohort and yielded a set of 795,063 SNPs after quality control (Le Clerc *et al.*, 2013). Genetic associations were searched with the severity of SL using linear regression after correcting for stratification and non-genetic confounding factors.

The description of the study population according to the severity grades of SL on cheeks and forehead is given in respectively table S1 and S2. The description of the population according to the facial SL score has already been described (Ezzedine *et al.*, 2013). The correlations between age and the three outcome variables are shown in Table 1 with the highest correlation found between SL assessed on the cheeks and globally on the face. As expected, age was significantly correlated with the three outcomes.

We first analysed the grade of SL on the forehead. Two genotyped SNPs, rs9350204 and rs9358294, in complete linkage disequilibrium (LD) in our sample, located on the chromosome 6, passed the genome-wide significance threshold (5×10^{-8}). These two SNPs were significant in the genotypic model with $P_{GENO} = 1.37 \times 10^{-8}$ and in the recessive model with $P_{REC} = 3.53 \times 10^{-8}$. The distribution of the genotype frequencies confirmed the recessive effect since all the homozygous rs9350204-CC individuals exhibited the most severe score (Figure S3). These two SNPs are located in an intergenic part of the 6p22 region. The nearest genes are the *RPL29P17* pseudogene and the *MBOAT1* gene at distances of approximately 50 kb and 100 kb respectively. A total of 18 distinct SNPs were in high LD ($r^2 > 0.8$) with at least one of the two significant SNPs, according to the European populations of the 1000 Genomes project. All these SNPs are also intergenic.

No genome-wide significant association was found for SL on the cheeks.

We then focused on the association between genotypes and the global score of facial SL and found 9 genotyped SNPs with a False Discovery Rate (FDR) below 25%, although none passed the genome-wide significance threshold (5×10^{-8}) (Table 2, Figures S4 and S5). Those SNPs gathered into two blocks located 11 Mb apart from each other in the chromosome 6. The first block was composed of the 2 SNPs (rs9350204 and rs9358294) in the 6p22 region (Figure 1) previously associated with the forehead SL (see above). In a similar fashion, homozygous rs9350204-CC individuals exhibited a higher global score of facial SL (Figure 2a and Figure S3). The second block, composed of 7 SNPs (rs2853949, rs2844614, rs2844613, rs2524069, rs2853947, rs2524067 and rs2524065), is located in the region 6p21 (Figure 3) and the heterozygous genotypes of these SNPs were correlated with more severe facial SL (Figure 2b). The SNPs were in high LD ($r^2 > 0.9$) within each block, but the two blocks did not exhibit any LD between each other (mean $r^2 = 0.03$) (Figure 4). The signals of each region were independent since p-values remained unchanged when adding SNPs of the other group as covariates in the tested model.

In the *HLA* 6p21 region, the rs2853949 SNP is located 2kb away from the 5'UTR of the *HLA-C* gene, and the remaining SNPs, except the intergenic SNP rs9266065, are located in the exonic region of the *USP8P1* pseudogene according to dbSNP (see Figure 3 and Table 2). According to the 1000 Genomes European sample, 16 SNPs were in high LD ($r^2 > 0.8$) with at least one of the 7 SNPs identified in this region. Among these 16 SNPs, 13 are intergenic. Two SNPs, rs2524065 and rs1049709, are respectively located in exonic regions of the *USP8P1* and *RPL3P2* pseudogenes and rs1049709 is in the 3'UTR part of the *HLA-C* gene (Figure 3). Interestingly, the SNPs identified in the 6p21 region play a role in the expression of *HLA* genes according to several eQTL databases (Table S6). For instance, the minor alleles

of this set of SNPs are significantly correlated with a decreased expression of *HLA-C* according to the three databases ($P_{genevar} = 5 \times 10^{-4}$, $P_{mRNA\_by\_SNP\_Browser} = 3 \times 10^{-7}$, $P_{GHS\_Express} = 6 \times 10^{-6}$).

Furthermore, we could impute 7 million SNP/indel variants from the 1000 Genomes project using bioinfomatics softwares and looked for associations with the three outcome variables. The imputation confirmed the associations found previously (region 6p21 and region 6p22) and no new signal passed neither the genome-wide significance threshold nor the FDR threshold. We were also able to impute the traditional class I and class II *HLA* alleles using the software SNP2HLA (Jia *et al.*, 2013). None of the *HLA* alleles was found significantly associated with SL. Yet, the *HLA-C*0701* allele exhibited a low p-value in the genotypic model when testing the associations with global facial SL, forehead and cheeks SL: $P_{GENO} = 9.70 \times 10^{-7}$, $P_{GENO} = 5.1 \times 10^{-7}$, $P_{GENO} = 9.82 \times 10^{-6}$, respectively. Strikingly, the *HLA-C*0701* allele was in high linkage disequilibrium with the top SNP in the 6p21 region (rs2524069-T allele) according to our genotype data ($r^2 = 0.95$).

Finally, we checked the p-values of the SNPs in 38 candidate genes previously found implicated in SL development by various studies. Interestingly, 19 SNPs passed the 25% FDR threshold, when testing associations with global SL on the face in the genotypic model. Among these 19 SNPs, 12 SNPs are located in *MITF* gene (best FDR = 0.13), two SNPs in *OCA2* gene (best FDR = 0.15), and one SNP for the *SCEL* (FDR = 0.13), *FABP7* (FDR = 0.20), *AMBP* (FDR = 0.20), and *RBBP6* (FDR = 0.24) genes (Table S7). None of them or those in linkage disequilibrium are located in coding regions, except for rs7593 SNP in *RBBP6* gene involving a synonymous change. However, rs62250968 and rs35017084 SNPs in high linkage disequilibrium ($r^2 > 0.8$) with rs9858495 and located in the *MITF* gene

are in two regions with high evidence for binding of transcription factors according to ENCODE.

Discussion

In this study, we have performed a GWAS looking for possible associations between SNPs and the severity of SL on the forehead, on the cheeks and globally on the face. Two groups of SNPs, located in regions 6p22 and 6p21, were found associated with a high score of SL globally on the face. In addition, the SNPs located in the 6p22 region were significantly associated with the severity of SL on the forehead. Imputation of the SNP/indel variants from the 1000 Genomes project did not reveal additional associations, but imputation of the *HLA* class I and II traditional alleles pointed to a role of the *HLA-C*0701* allele.

On the one hand, the SNPs located in the 6p22 block are intergenic and we could not derive any relevant biological interpretation. On the other hand, one SNP in the 6p21 region is located in the 5'UTR region of the *HLA-C* gene and the set of SNPs belonging to this block is significantly associated with the levels of expression of several *HLA* genes, notably *HLA-C*. Moreover, the imputation of the *HLA* alleles revealed for *HLA-C*0701* a trend for association with forehead SL ($P_{GENO} = 9.70 \times 10^{-7}$), and we found that this latter *HLA-C* allele is in high LD with the 6p21 associated SNPs ($r^2 = 0.95$). The *HLA-C*0701* allele is associated with a decreased expression of HLA-C molecules and with a higher susceptibility to facial SL in the dominant mode. HLA-C belongs to the HLA class I molecules and consists of a heavy chain and a light chain (beta-2 microglobulin). HLA-C plays a dual role in immunity in that it can present antigens to cytotoxic T lymphocytes (CTLs) and can inhibit natural killer cells through interaction with the KIR receptors (Blais *et al.*, 2011). Overall, our results argue for an association between a lower *HLA-C* expression and the severity of SL. A contrario, high expression of *HLA-C* has been associated with the control of β-HPV induced squamous cell carcinoma (Vineretsky *et al.*, 2014). These observations could suggest a mechanism of control

of abnormal skin cells through *HLA-C* recognition. In such a scenario aberrant melanocytes would be not eliminated by surveying immune cells.

Our study was based on the use of photographs to assess the severity of SL. This may explain some discrepancies among the associations found for the three outcomes. However, these photographs were standardized, limiting possible misinterpretations. eQTL studies on the *HLA* locus must also be interpreted with caution because the high degree of genetic variability and linkage disequilibrium across the *HLA* region (Cookson *et al.*, 2009). As for any GWAS, it will thus be important to replicate these results in independent cohorts.

We have also investigated associations with SNPs from 38 candidate genes, including several genes involved in the inflammation in SL cells (Aoki *et al.*, 2007). We found 19 SNPs passing 25% FDR threshold, with the strongest signals located in the *MITF* gene. *MITF* encodes a key transcription factor for melanogenesis. It is involved in the differentiation, growth and survival of pigment cells, by regulating melanocyte-specific transcription of melanogenesis-related enzyme genes (Tachibana, 2000; Yasumoto *et al.*, 1997). According to the Encode project, two SNPs rs62250968 and rs35017084 located in *MITF* are likely to affect binding of transcription factors and should thus affect *MITF* gene expression.

Our results underline the complexity of the mechanisms governing the pathogenesis of SL, which involve HLA/immunity, inflammation, and melanin metabolism. As next steps following our GWA study, it will be of interest to use systems biology approaches by integrating gene sets of pathway analysis for the molecular etiology of the development of SL.

Materials and methods

Study design and population

This population was previously described in details (Le Clerc *et al.*, 2013). Briefly, in the autumn/winter of 2002-2003, 570 of the 2,257 middle-aged women living in the Paris area from the SU.VI.MAX cohort (Hercberg *et al.*, 1998) agreed to participate in a research on skin aging and provided informed consent. Each participant completed a self-administered questionnaire related to lifetime sun exposure behavior and three standardized high-resolution digital images (2,008 x 3,032 pixels) of the face were taken under normalized lightning conditions (one frontal view of the face and one of each profile), using a Kodak DCS 760 digital camera with a 105 mm camera lens (Kodak, Paris, France).

Outcome variables: phenotype analyzed

After images acquisition, solar lentigines were evaluated separately on both cheeks and forehead by a dermatologist using a specific six-grade scale with photographic illustrations (Figure S8). Then, the severity of facial solar lentigine was estimated by a score based on Principal Component Analysis and linear regression (Ezzedine *et al.*, 2013). Both of the two area-specific grades and the facial score were analysed in the study.

Covariates used for the statistical analysis

Several characteristics susceptible to play a role on the onset of solar lentigines (Ezzedine *et al.*, 2013) were taken into account: age (in years), body mass index (BMI; in $kg.m^{-2}$), smoking habits (never, former and current), and hormonal status (nonmenopausal, menopausal with

hormone replacement therapy, and menopausal without hormone replacement therapy). BMI was categorized as underweight or normal (BMI < 25 $kg.m^{-2}$), overweight (25 ≤ BMI < 30 $kg.m^{-2}$), or obese (BMI ≥ 30 $kg.m^{-2}$) according to the World Health Organization (WHO) recommendations (WHO, 1995). Moreover, lifetime sun exposure intensity was estimated by a score based on data collected by a self-reported questionnaire (Ezzedine *et al.*, 2013). The design, validation, and description of this score have been described previously (Guinot *et al*, 2001).

Genotyping method

The genotyping method has already been described in details (Le Clerc *et al.*, 2013). The 529 women were genotyped using Illumina Infinium HumanOmni1-Quad BeadChips (Illumina, San Diego, CA) that contain 1,140,419 markers and a sample of 250 ng of ADN by individual was used to obtain genotypes. For the analysis, we considered only SNPs, consequently excluding the copy-number variations that represented 91,706 markers on the HumanOmni1-Quad BeadChips. Moreover, 2,182 SNPs located on the Y chromosome were removed.

Quality control

The quality control steps have already been described in a previous study (Le Clerc *et al.*, 2013). Briefly, 9 samples with a call rate (percentage of SNPs genotyped by sample) of < 95% in the Illumina clusters were removed. The SNPs with a call frequency (percentage of samples genotyped by SNP) of < 99% were reclustered and, after that, samples with a call rate of < 98% were deleted. In total, after these quality control steps, 56,479 SNPs with a call

frequency of < 98% (2% of missing data) were excluded. Hardy–Weinberg equilibrium analysis was performed by using an exact statistical test implemented in PLINK software (Purcell *et al.*, 2007) was performed for each SNP. Thus, 3,866 SNPs, which were not in the Hardy–Weinberg equilibrium ($P < 1 \times 10^{-3}$), were rejected. Finally, we removed 191,123 SNPs with minor allele frequency < 1% to avoid error of genotyping, leaving a total of 795,063 SNPs.

Identification of population stratification

To correct for possible population stratification, genotypes were analyzed using EIGENSTRAT utility of the EIGENSOFT package version 4.2 (Price *et al.*, 2006). The two first pass with the Eigenstrat software pointed out 18 outliers, who were removed from further analyses. Then, a third pass without outliers was performed to determine the Eigen vectors. In the statistical analysis, we used the top two Eigen vectors as covariates to correct for population substructure in the association analyses.

Statistical analysis

The population was first described according to the severity of each of the grades of lentigines on the cheeks and on the forehead, using a series of analyses of variance for quantitative variables and using χ^2 tests for qualitative variables. Then, the associations between the genotypes and each of the three outcomes were tested.

The statistical analysis was performed using a multivariate linear regression implemented in the PLINK software (Purcell *et al.*, 2007) in the additive, dominant recessive and genotypic

model, taking as covariates the two first Eigenstrat principal components and the potential confounding factors (smoking habits, BMI, hormonal status, lifetime sun exposure intensity, and age). The genome-wide significance threshold (5×10^{-8}) or a False Discovery Rate (Benjamini and Hochberg, 1995) threshold of 25% were used to assess statistical significance. The q-values were computed using the R package 'qvalues' (available at http://www.bioconductor.org/packages/release/bioc/html/qvalue.html). Finally, the imputed data were similarly analyzed using the SNPTEST software Only SNPs with imputation quality score > 0.8, MAF > 1% and test score > 0.9 were kept.

Linkage disequilibrium

For each SNP exhibiting a significant association, we looked for SNPs in high linkage disequilibrium ($r^2 > 0.8$) in the European populations of 1000 Genomes (EUR, 1000 Genomes Phase I integrated release version 3, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/). We retained only the set of distinct SNPs in high LD with at least one significantly associated SNP.

Imputation for high density mapping and for *HLA* alleles

The genotype data were phased using SHAPEIT2 (Delaneau *et al.*, 2013) The phased data were then imputed using IMPUTE2 (Howie *et al.*, 2009). As reference haplotypes, we used genotype data of 1094 individuals from the phase I integrated variant set of the 1000 Genomes project released in March 2012 and updated in August 2012 (Abecasis *et al.*, 2010). To reduce uncertainty, SNPs with MAF < 1% in European samples of the reference panel

were not used to impute our data. The imputation of *HLA* alleles was done using SNP2HLA (Jia *et al.*, 2013).

Exploration of genes already identified in the literature

To compare our results to those already known, we checked the p-values of 38 genes identified by former studies (Aoki *et al.*, 2007; Bastiaens *et al.*, 2001; Cardinali *et al.*, 2012; Hafner *et al.*, 2009a; Hafner *et al.*, 2009b; Vierkotter *et al.*, 2012): GENE (gene ID); *ACSBG1* (23205); *AMBP* (259); *BAMBI* (25805); *CIDEA* (1149); *DCT* (1638); *EDN1* (1906); *ELOVL3* (83401); *FA2H* (79152); *FABP7* (2173); *FADS1* (3992); *FADS2* (9415); *FAR2* (55711); *FGFR* (32261); *GAL* (51083); *HAO2* (51179); *HSD3B1* (3283); *INSIG1* (3638); *KIT*, (3815); *KITLG* (4254); *LZTS1* (11178); *MIA* (8190); *MITF* (4286); *MLANA* (2315); *MOGAT1* (116255); *MSMB*, (4477); *NKTR* (4820); *OCA2* (4948); *PECR* (55825); *PMEL* (6490); *POMC* (5443); *PRPF38B* (55119); *RBBP6* (5930); *SCEL* (8796); *SEC14L4* (284904); *TRPM1* (4308); *TYR* (7299); *TYRP1* (7306); *UBIAD1* (29914). The *MC1R* gene, which has already been associated with SL (Bastiaens *et al.*, 2001), has previously been investigated in our cohort and was involved in freckles but not in lentigines (Ezzedine *et al.*, 2013). For this reason, *MITF* was not included in this list.

To obtain a single p-value for each gene, we identified all the either genotyped or imputed SNPs in our data located in this particular gene. To take regulatory regions into account, genes were enlarged 5 kb downstream and upstream. We assessed the lowest single SNP p-value to this gene.

Bioinformatics exploration

To further explore the signals from GWAS, we peformed in silico searches for possible SNP alterations, using several databases. Gene expression: Genevar (http://www.sanger.ac.uk/resources/software/genevar/), mRNA by SNP Browser http://www.sph.umich.edu/csg/liang/asthma/), and GHS-Express (http://genecanvas.ecgene.net/uploads/ForReview/); methylation: Genevar (http://www.sanger.ac.uk/resources/software/genevar/); polyadenylation regions: PolyApred (http://www.imtech.res.in/raghava/polyapred/submission.html); transcription factor binding sites: regulomedb (http://regulomedb.org/ derived from TRANSFAC database); miRNA genes or miRNA targets: miRBAse (http://www.mirbase.org/), miRTarBase (http://mirtarbase.mbc.nctu.edu.tw) and MicroCosm Targets (http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/); ENCODE data: regulomedb (http://regulomedb.org/).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

V. L. and S.L.C contributed equally to this work. The authors gratefully acknowledge the dedicated efforts of all the SU.VI.MAX volunteers, the investigators, and the staff members involved in this study, especially Dr Sandrine Bertrais, and Ms Nathalie Arnault and Mr Gwenael Monot who coordinated the data management.

Bibliography

Abecasis GR, Altshuler D, Auton A*, et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061-73.

Andersen WK, Labadie RR, Bhawan J (1997) Histopathology of solar lentigines of the face: a quantitative study. *J Am Acad Dermatol* 36:444-7.

Aoki H, Moro O, Tagami H*, et al.* (2007) Gene expression profiling analysis of solar lentigo in relation to immunohistochemical characteristics. *Br J Dermatol* 156:1214-23.

Bastiaens M, Hoefnagel J, Westendorp R*, et al.* (2004) Solar lentigines are strongly related to sun exposure in contrast to ephelides. *Pigment Cell Res* 17:225-9.

Bastiaens M, ter Huurne J, Gruis N*, et al.* (2001) The melanocortin-1-receptor gene is the major freckle gene. *Hum Mol Genet* 10:1701-8.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 57:289-300.

Blais ME, Dong T, Rowland-Jones S (2011) HLA-C as a mediator of natural killer and T-cell activation: spectator or key player? *Immunology* 133:1-7.

Cardinali G, Kovacs D, Picardo M (2012) Mechanisms underlying post-inflammatory hyperpigmentation: lessons from solar lentigo. *Ann Dermatol Venereol* 139 Suppl 4:S148-52.

Chang AL, Atzmon G, Bergman A*, et al.* (2014) Identification of genes promoting skin youthfulness by genome-wide association study. *J Invest Dermatol* 134:651-7.

Cookson W, Liang L, Abecasis G*, et al.* (2009) Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10:184-94.

Delaneau O, Zagury JF, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10:5-6.

Elfakir A, Ezzedine K, Latreille J*, et al.* (2010) Functional MC1R-gene variants are associated with increased risk for severe photoaging of facial skin. *J Invest Dermatol* 130:1107-15.

Ezzedine K, Mauger E, Latreille J*, et al.* (2013) Freckles and solar lentigines have different risk factors in Caucasian women. *J Eur Acad Dermatol Venereol* 27:e345-56.

Guinot C, Malvy D, Latreille J *et al*. (2001). Sun exposure behaviour of a general adult population in France. In: Ring J, Weidinger S, Darsow U eds *Skin and Environment - Perception and Protection*. Monduzzi editore S.p.A: Bologna, 1099–106

Hafner C, Stoehr R, van Oers JM*, et al.* (2009a) The absence of BRAF, FGFR3, and PIK3CA mutations differentiates lentigo simplex from melanocytic nevus and solar lentigo. *J Invest Dermatol* 129:2730-5.

Hafner C, Stoehr R, van Oers JM*, et al.* (2009b) FGFR3 and PIK3CA mutations are involved in the molecular pathogenesis of solar lentigo. *Br J Dermatol* 160:546-51.

Hercberg S, Galan P, Preziosi P*, et al.* (1998) Background and rationale behind the SU.VI.MAX Study, a prevention trial using nutritional doses of a combination of antioxidant vitamins and minerals to reduce cardiovascular diseases and cancers. SUpplementation en VItamines et Mineraux AntioXydants Study. *Int J Vitam Nutr Res* 68:3-20.

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5:e1000529.

Jia X, Han B, Onengut-Gumuscu S*, et al.* (2013) Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* 8:e64683.

Le Clerc S, Taing L, Ezzedine K*, et al.* (2013) A genome-wide association study in Caucasian women points out a putative role of the STXBP5L gene in facial photoaging. *J Invest Dermatol* 133:929-35.

Price AL, Patterson NJ, Plenge RM*, et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904-9.

Purcell S, Neale B, Todd-Brown K*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559-75.

Tachibana M (2000) MITF: a stream flowing for pigment cells. *Pigment Cell Res* 13:230-40.

Vierkotter A, Kramer U, Sugiri D*, et al.* (2012) Development of lentigines in German and Japanese women correlates with variants in the SLC45A2 gene. *J Invest Dermatol* 132:733-6.

Vineretsky KA, Karagas MR, Kuriger-Laber JK*, et al.* (2014) HLA-C -35kb expression SNP is associated with differential control of beta-HPV infection in squamous cell carcinoma cases and controls. *PLoS One* 9:e103710.

Yaar M, Gilchrest BA (2007) Photoageing: mechanism, prevention and therapy. *Br J Dermatol* 157:874-87.

Yasumoto K, Yokoyama K, Takahashi K*, et al.* (1997) Functional analysis of microphthalmia-associated transcription factor in pigment cell-specific transcription of the human tyrosinase family genes. *J Biol Chem* 272:503-9.

Tables

Table 1. Correlation coefficients between age and the three outcome variables

	Age	Grade of solar lentigine on the forehead	Grade of solar lentigine on the cheeks	Global score of facial solar lentigine
Age	1	0.20*	0.18*	0.19*
Grade of solar lentigine on the forehead		1	0.51*	0.70*
Grade of solar lentigine on the cheeks			1	0.85*
Score of facial solar lentigine				1

*: $P < 0.0001$

Table 2. SNPs associated with the global score of facial solar lentigines (FDR < 0.25)

SNP	Region	Position	Minor Allele	MAF	P-value	Q-value	Localisation
rs9350204	6p22	19996808	C	0.154	1.61×10^{-6}	0.129	Intergenic
rs9358294	6p22	19999122	G	0.154	1.61×10^{-6}	0.129	Intergenic
rs2853949	6p21	31241664	A	0.136	6.24×10^{-7}	0.084	*HLA-C* (Near Gene)
rs2844614	6p21	31243540	A	0.136	4.89×10^{-7}	0.084	*USP8P1* (exon)
rs2844613	6p21	31243846	T	0.136	6.95×10^{-7}	0.084	*USP8P1* (exon)
rs2524069	6p21	31244789	T	0.130	2.53×10^{-7}	0.084	*USP8P1* (exon)
rs2853947	6p21	31245796	T	0.136	5.57×10^{-7}	0.084	*USP8P1* (exon)
rs2524067	6p21	31245821	G	0.136	1.12×10^{-6}	0.115	*USP8P1* (exon)
rs2524065	6p21	31246324	C	0.136	4.89×10^{-7}	0.084	*USP8P1* (exon)

MAF, Minor Allele Frequency

Figure Legends

Figure 1. Association results for the facial solar lentigine score and genetic map of the 6p22 region.

At the top, association results of genotyped and imputed SNPs in the locus 6p22 are plotted as the distribution of the -log10(P) along the physical position. The top rs9350204 SNP is represented in purple and all the others SNPs are colored in function of their degree of linkage disequilibrium with rs9350204, according to the r^2 legend at the right top. At the bottom, the genetic map of the region covered by the top SNPs. The significant genotyped SNPs (purple and red) and not genotyped SNPs in high linkage disequilibrium (black, $r^2 > 0.8$) are represented. There are no exonic SNPs. Grey boxes represent exons (genes *ID4* and *MBOAT1*) and black boxes represent the pseudogene *RPL29P17*. The arrows indicate the direction of transcription.

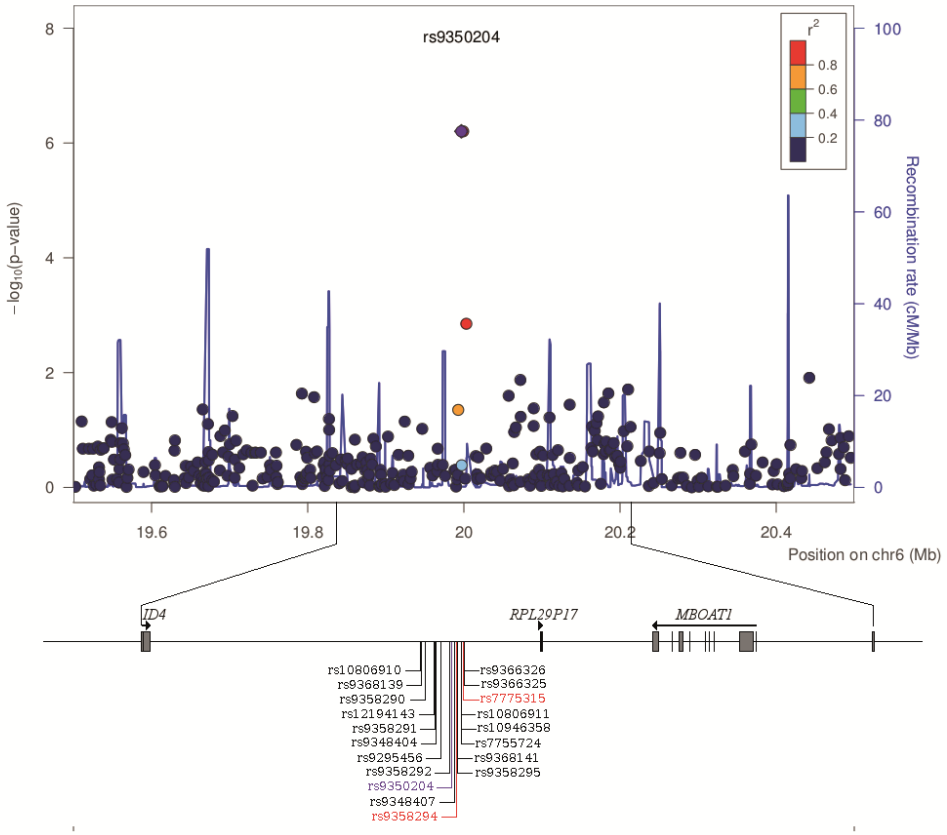


Figure 2. Distribution of the score of facial solar lentigines in respect to the different genotypes.

a) Distribution of the score of facial solar lentigines in respect to the genotype of the SNP rs9350204. b) Distribution of the score of facial solar lentigines in respect to the genotype of the SNP rs2524069.

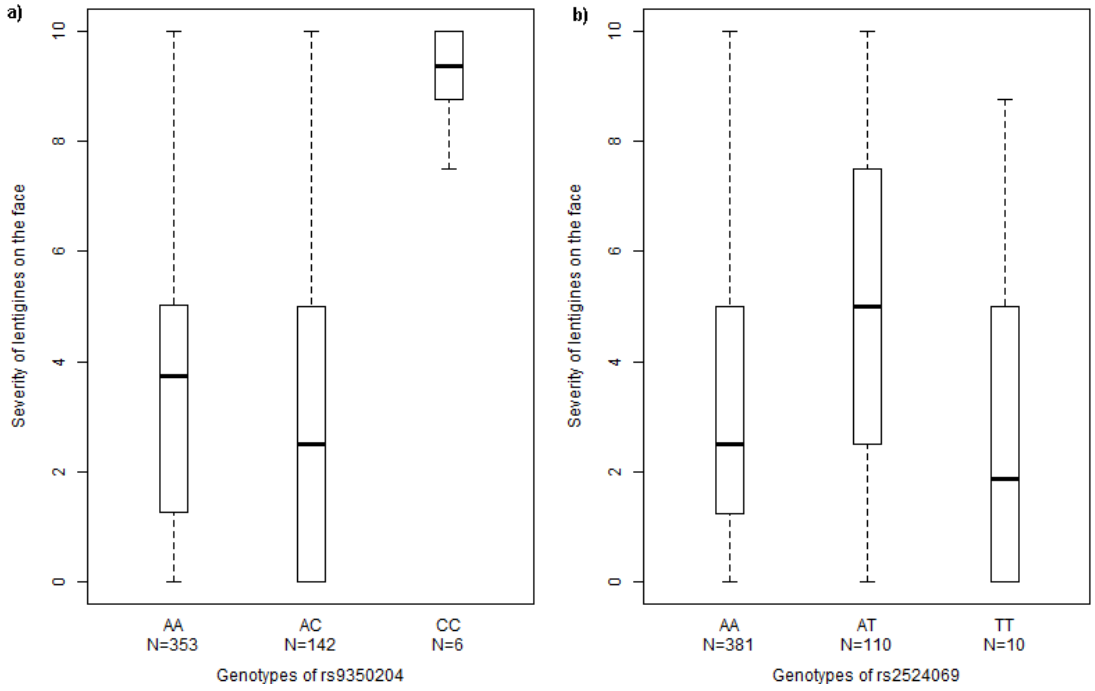


Figure 3. Association results for the facial solar lentigine score and genetic map of the 6p21 region.

At the top, association results of genotyped and imputed SNPs in the locus 6p21 are plotted as the distribution of the -log10(P) along the physical position. The top rs2844614 SNP is represented in purple and all the others SNPs are colored in function of their degree of linkage disequilibrium with rs2844614, according to the r^2 legend at the right top. At the bottom, the genetic map of the region covered by the best SNPs. The most significant genotyped SNP is in purple, the genotyped SNPs with $r^2 > 0.8$ are in red, and SNPs in high linkage disequilibrium ($r^2 > 0.8$) but non genotyped are in black. Exonic SNPs are flagged with an asterisk (*). Grey boxes represent exons (genes *HLA-C* and *HLA-B*) and black boxes represent the pseudogenes (*USP8P1* and *RPL3P2*). The arrows indicate the direction of transcription.

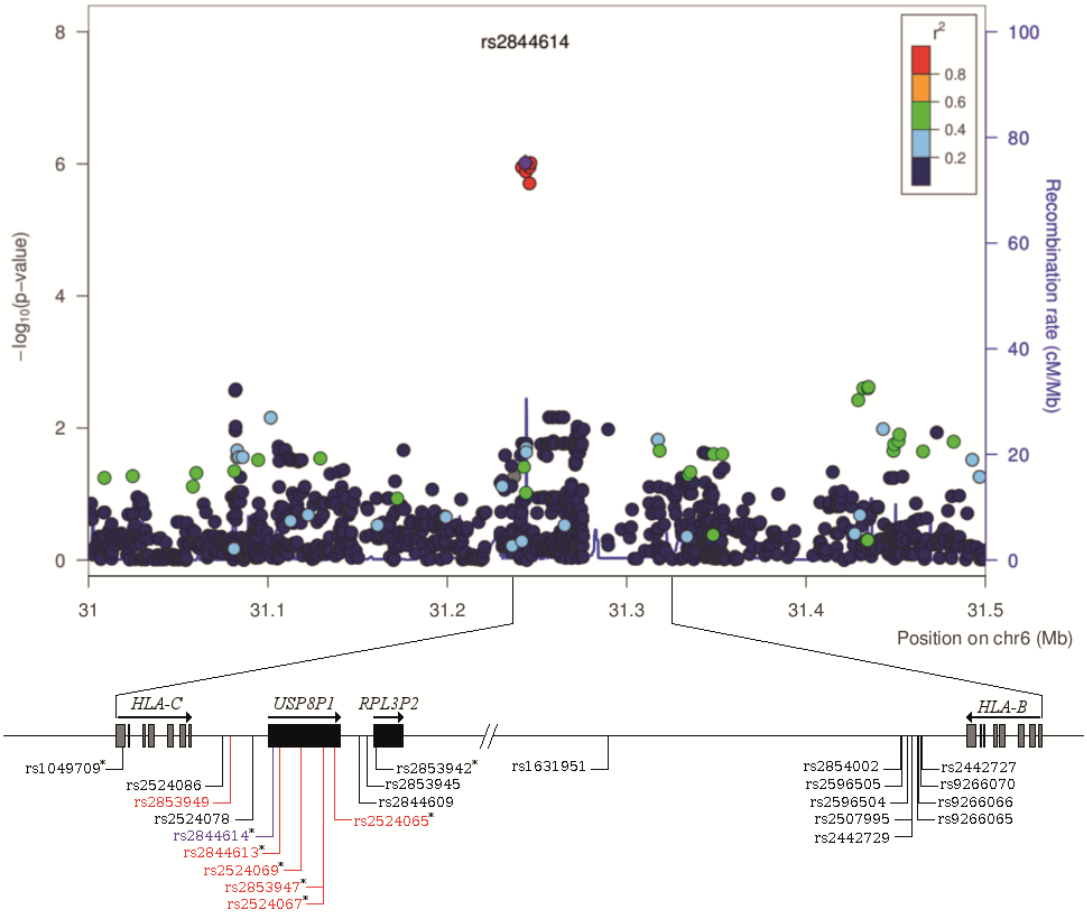
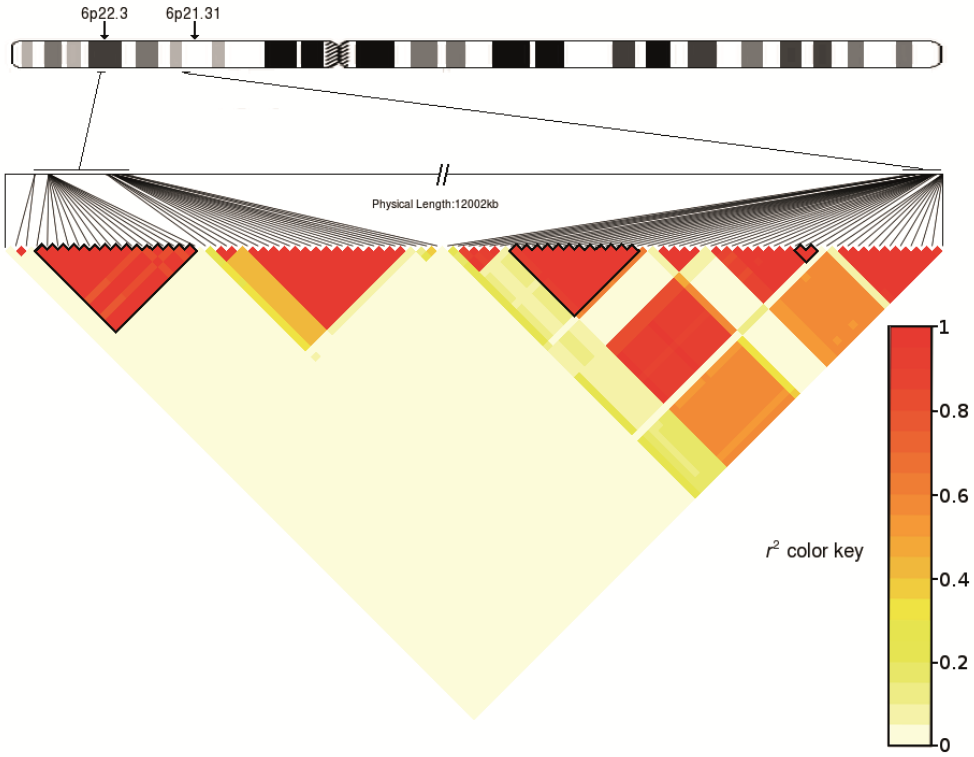


Figure 4. Linkage disequilibrium map of SNPs in the 6p22 and 6p21 regions, exhibiting p-values under 10^{-3} for association with the score of facial solar lentigines.

The LD map shows the r^2 coefficient between SNPs with p-value under 10^{-3} in the 6p22 and 6p21 regions. SNPs identified in the association study are highlighted in black triangles. All SNPs in each region are in high LD but the r^2 coefficients between the two regions are very low.



Supplemental data

A Genome-Wide Association Study in Caucasian Women reveals the involvement of HLA genes in the severity of facial solar lentigines

Supplementary Table S1. Description of the population according to lentigines severity on the cheeks.

	Severity of age points on the cheeks						
	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4/5[1]	Total	P-value of
	N=120	N=127	N=131	N=77	N=46	N=501	test
Age (years)	55.9±6.8[2]	56.6±6.1	58.1±6.3	59.5±6.0	60.0±5.6	57.6±6.4	<0.0001[4]
Lifetime sun exposure (score)	4.5±3.5	5.3±3.5	5.4±3.6	5.6±3.6	6.5±3.4	5.3±3.5	0.0008[3]
BMI classification							0.61[5]
Normal	80 (16.0)[5]	87 (17.4)	83 (16.6)	51 (10.2)	33 (6.6)	334 (66.7)	
Overweight	30 (6.0)	24 (4.8)	36 (7.2)	21 (4.2)	11 (2.2)	122 (22.4)	
Obese	10 (2.0)	16 (3.2)	12 (2.4)	5 (1.0)	2 (0.4)	45 (9.0)	
Hormonal status							0.22[4]
No menopausal	32 (6.4)	24 (4.8)	25 (5.0)	8 (1.6)	5 (1.0)	94 (18.8)	
Menopausal with HRT	57 (11.4)	67 (13.4)	66 (13.2)	44 (8.9)	27 (5.4)	261 (52.1)	
Menopausal without HRT	31 (6.2)	36 (7.2)	40 (8.0)	25 (5.0)	14 (3.0)	146 (29.1)	
Smoking habits							0.04[4]
Never	56 (11.2)	74 (14.8)	76 (15.2)	48 (9.6)	32 (6.4)	286 (57.1)	
Former smoker	51 (10.2)	39 (7.8)	34 (6.8)	24 (4.8)	12 (2.4)	160 (31.9)	
Current smoker	13 (2.6)	14 (2.8)	21 (4.2)	5 (1.0)	2 (0.4)	55 (11.0)	
Eye color							0.02[4]
Blue/grey	40 (8.0)	41 (8.2)	33 (6.6)	17 (3.4)	5 (1.0)	136 (27.3)	
Green/hazel/brown/black	79 (15.8)	85 (17.0)	98 (19.6)	60 (12.0)	41 (8.2)	363 (72.7)	
Hair color at twenty							0.07[4]
Blond/red	35 (7.0)	27 (5.4)	29 (5.8)	10 (2.0)	7 (1.4)	108 (21.6)	
Light & dark brown/black	84 (16.8)	99 (19.8)	102 (20.4)	67 (13.4)	39 (7.8)	391 (78.4)	
Skin color without tanning							0.17[4]
Fair	100 (20.0)	96 (19.2)	97 (19.4)	63 (12.6)	32 (6.4)	388 (77.8)	
Dark	19 (3.8)	30 (6.0)	34 (6.8)	14 (2.8)	14 (2.8)	111 (22.2)	
History of facial freckles							< 0.0001[4]
No	83 (16.6)	90 (18.0)	66 (13.2)	38 (7.6)	17 (3.4)	294 (58.9)	
Yes	36 (7.2)	36 (7.2)	65 (13.0)	39 (7.8)	29 (5.8)	205 (41.1)	
Suntan intensity							0.01[4]
None/slight/light	86 (17.2)	72 (14.4)	70 (14.0)	49 (9.8)	23 (4.6)	300 (60.1)	
Dark/very dark	33 (6.6)	54 (10.8)	61 (12.2)	28 (5.6)	23 (4.6)	199 (39.9)	
Sunburn event frequency							0.45[4]
None/rare	74 (14.8)	95 (19.0)	98 (19.6)	55 (19.6)	36 (7.2)	358 (71.7)	
Frequent/constant	45 (9.0)	31 (6.2)	33 (6.6)	22 (4.4)	10 (2.0)	141 (28.3)	

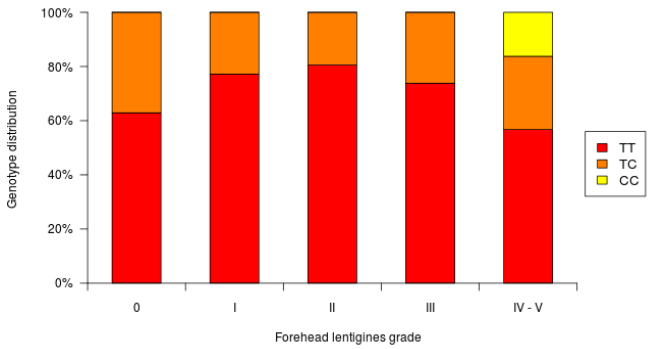
[1] As a single woman had grade-6, she had been grouped with grade-5 individuals.
[2] Mean ± Standard Deviation
[3] Frequency and (%), due to possible missing values the sum of the cell frequencies can be smaller than the total indicated in the top of the columns.
[4] ANOVA test
[5] χ^2 test

Supplementary table S2. Description of the population according to lentigines severity on the forehead.

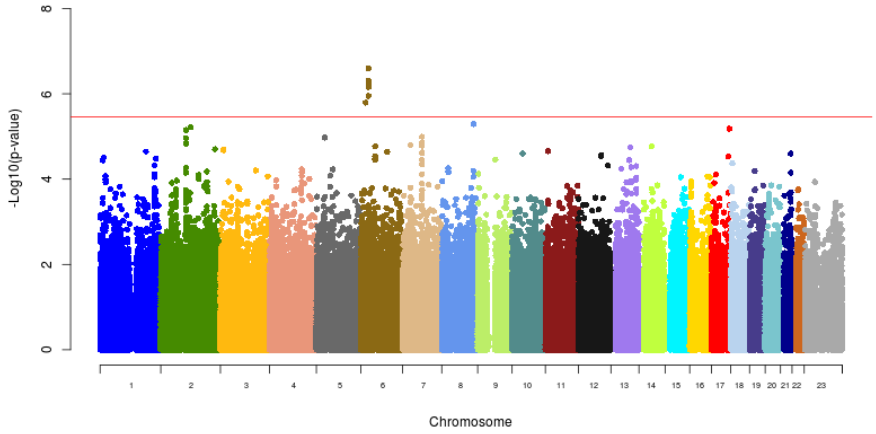
	Severity of age points on the forehead						
	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4/5[1]	Total	P value of test
	N=186	N=136	N=77	N=65	N=37	N=501	
Age (years)	56.0±6.6[2]	57.2±5.7	58.5±6.5	60.3±5.8	60.7±5.7	57.6±6.4	< 0.0001[3]
Lifetime sun exposure (score)	4.7±3.5	5.1±3.5	5.8±3.3	6.0±3.6	6.5±3.6	5.3±3.5	0.01[3]
BMI classification							0.75[4]
Normal	123	96 (19.2)	52 (10.4)	41 (8.2)	22 (4.4)	334 (66.7)	
Overweight	46 (9.2)	29 (5.8)	16 (3.2)	18 (3.6)	13 (2.6)	122 (22.4)	
Obese	17 (3.4)	11 (2.2)	9 (1.8)	6 (1.2)	2 (0.4)	45 (9.0)	
Hormonal status							0.20[4]
No menopausal	45 (9.0)	25 (5.0)	14 (2.8)	7 (1.4)	3 (0.6)	94 (18.8)	
Menopausal with HRT	90 (18.0)	69 (13.8)	39 (7.8)	38 (7.6)	25 (5.0)	261 (52.1)	
Menopausal without HRT	51 (10.2)	42 (8.4)	24 (4.8)	20 (4.0)	9 (1.8)	146 (29.1)	
Smoking habits							0.94[4]
Never	106 (21.2)	76 (15.2)	42 (8.4)	40 (8.0)	22 (4.4)	286 (57.1)	
Former smoker	58 (11.6)	48 (9.6)	25 (5.0)	17 (3.4)	12 (2.4)	160 (31.9)	
Current smoker	22 (4.4)	12 (2.4)	10 (2.0)	8 (1.6)	3 (0.6)	55 (11.0)	
Eye color							0.001[4]
Blue/grey	60 (12.0)	47 (9.4)	13 (2.6)	8 (1.6)	8 (1.6)	136 (27.3)	
Green/hazel/brown/black	124 (24.8)	89 (17.8)	64 (12.8)	57 (11.4)	29 (5.8)	363 (72.7)	
Hair color at twenty							0.14[4]
Blond/red	47 (9.4)	24 (4.8)	11 (2.2)	15 (3.0)	11 (2.2)	108 (21.6)	
Light&dark brown/black	137 (27.5)	112 (22.4)	66 (13.2)	50 (10.0)	26 (5.2)	391 (78.4)	
Skin color without tanning							0.97[4]
Fair	145 (29.1)	106 (21.2)	58 (11.6)	51 (10.2)	28 (5.6)	388 (77.8)	
Dark	39 (7.8)	30 (6.0)	19 (3.8)	14 (2.8)	9 (1.8)	111 (22.2)	
History of facial freckles							< 0.0001[4]
No	128 (25.7)	84 (16.8)	40 (8.0)	31 (6.2)	11 (2.2)	294 (58.9)	
Yes	56 (11.2)	52 (10.4)	37 (7.4)	34 (6.8)	26 (5.2)	205 (41.1)	
Suntan intensity							0.07[4]
None/slight/light	125 (25.1)	79 (15.8)	41 (8.2)	33 (6.6)	22 (4.4)	300 (60.1)	
Dark/very dark	59 (11.8)	57 (11.4)	36 (7.2)	32 (6.4)	15 (3.0)	199 (39.9)	
Sunburn event frequency							0.45[4]
None/rare	131 (26.3)	101 (20.2)	55 (11.0)	49 (9.8)	22 (4.4)	358 (71.7)	
Frequent/constant	53 (10.6)	35 (7.0)	22 (4.4)	16 (3.2)	15 (3.0)	141 (28.3)	

[1] As a single woman had grade-6, she had been grouped with grade-5 individuals.
[2] Mean ± Standard Deviation
[3] Frequency and (%), due to possible missing values the sum of the cell frequencies can be smaller than the total indicated in the top of the columns.
[4] ANOVA test
[5] χ^2 test

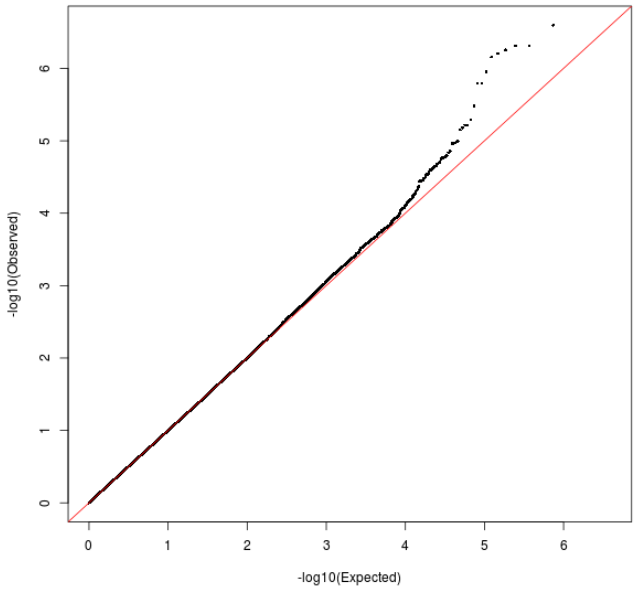
Supplementary figure S3. Distribution of the rs9350204 genotypes in function of the lentigines on the forehead grade severity.



Supplementary figure S4. Manhattan plot of the association study with the global score of facial lentigines.



Supplementary figure S5. Quantile-quantile plot of the association study with the global score of facial lentigines.



Supplementary table S6. Genes which expression is impacted by SNPs identified in the 6p21 region, according to three eQTL databases.

Genes	Genevar	mRNA by SNP Browser v 1.0.1	GHS Express
HLA-C	$5x10^{-4}$	$4x10^{-7}$	$6x10^{-6}$
BTN3A2	-	-	$9x10^{-26}$
HLA-A	-	-	$5x10^{-12}$
HLA-B	-	-	$4x10^{-6}$
HLA-DPB1	-	-	$4x10^{-16}$
HLA-DQA1	-	$4x10^{-13}$	$3x10^{-15}$
HLA-DQB1	-	$2x10^{-8}$	$8x10^{-16}$
HLA-DRB1	-	$2x10^{-7}$	$6x10^{-21}$
HLA-DRB3	-	-	$9x10^{-9}$
HLA-DRB4	-	$2x10^{-5}$	$7x10^{-40}$

Supplementary table S7. SNPs passing the 25% FDR threshold in the candidate genes analysis.

SNP	chr	position	gene	type	P-value	Q-value
rs17006579	3	70017219	*MITF*	intron	$6.1x10^{-4}$	0.13
rs7430957	3	70061156	*MITF*	intron	$8.7x10^{-4}$	0.13
rs9858495	3	69956708	*MITF*	intron	$1.3x10^{-3}$	0.13
rs1529585	3	69953285	*MITF*	intron	$1.3x10^{-3}$	0.13
rs1121878	3	69880016	*MITF*	intron	$1.5x10^{-3}$	0.13
rs576	3	70099157	*MITF*	3'UTR	$1.8x10^{-3}$	0.13
rs1491687	3	70100391	*MITF*	3'	$1.8x10^{-3}$	0.13
rs6777363	3	69926661	*MITF*	intron	$1.8x10^{-3}$	0.13
rs704246	3	70099594	*MITF*	3'	$2.0x10^{-3}$	0.13
rs13071701	3	69985204	*MITF*	intron	$2.3x10^{-3}$	0.13
rs7034903	9	115865496	*AMBP*	intron	$2.3x10^{-3}$	0.13
rs912292	13	77093535	*SCEL*	intron	$2.3x10^{-3}$	0.13
rs17650960	15	25673037	*OCA2*	3'	$2.8x10^{-3}$	0.15
rs7623610	3	70087971	*MITF*	intron	$3.2x10^{-3}$	0.16
rs1498519	15	25685246	*OCA2*	intron	$3.5x10^{-3}$	0.17
rs6904500	6	123151295	*FABP7*	3'	$4.5x10^{-3}$	0.20
rs1544978	3	69982827	*MITF*	intron	$5.1x10^{-3}$	0.20
rs12377342	9	115870140	*AMBP*	intron	$5.2x10^{-3}$	0.20
rs7593	16	24490907	*RBBP6*	exon syn	$6.4x10^{-3}$	0.24

The column "type" specifies the nature of the variant: "intron" for the intronic SNPs, "exon syn" for synonymous exonic SNPs, and "3'" means that the SNP is located at the 3' side next the gene. Abbreviation: chr, chromosome; syn, synonymous

Supplementary Figure S8. Photographic scale used to evaluate lentigines severtiy on the cheeks.

2 Etude “génome entier” sur une cohorte de femmes caucasiennes à la recherche de gènes associés à l’affaissement de la paupière

A Genome-Wide Association Study identifies new genetic associations in upper eyelid sagging.

Vincent Laville*, Sigrid Le Clerc*, Khaled Ezzedine, Randa Jdid, Lieng Taing, Toufik Labib, Cedric Coulonges, Damien Ulveling, Wassila Carpentier, Pilar Galan, Serge Hercberg, Frederique Morizot, Julie Latreille, Denis Malvy, Erwin Tschachler, Christiane Guinot, Jean-François Zagury

Papier en préparation

Pour tenter d’élucider de façon la plus exhaustive possible les facteurs génétiques impliqués dans l’affaissement de la paupière supérieure, nous avons réalisé une étude d’association génome-entier sur 520 femmes françaises et d’âge moyen issues de la cohorte SU.VI.MAX. La sévérité de l’affaissement de la paupière a été mesurée à l’aide d’un grade évalué à partir de photographies standardisées [176].

Nos résultats nous ont permis d’identifier deux SNPs sur le chromosome 10 et un SNP sur le chromosome 15 significativement associés à la sévérité de l’affaissement de la paupière supérieure.

Sur le chromosome 10, deux SNPs en déséquilibre de liaison total, rs4746957 ($P = 1,69 \times 10^{8}$) et rs16927253 ($P = 2,23 \times 10^{10}$), avaient une p-valeur d’association dans le mode génotypique inférieure au seuil de significativité génome-entier (5×10^{8}). Ces deux SNPs sont situés dans un intron du gène *H2AFY2*, codant pour une protéine membre de la famille des histones. L’imputation de nos données génotypées a permis la mise en évidence de 10 SNPs et un indel, en fort LD avec les SNPs génotypés identifiés ($r^2 > 0,8$), significativement associés au

phénotype. De plus, des SNPs proches du gène *COL13A1*, codant pour le collagène de type XIII, avaient des p-valeurs faibles suggérant ainsi un possible lien entre ce gène et la sévérité de l'affaissement de la paupière.

Sur le chromosome 15, rs4842897 ($P = 9,13 \times 10^{12}$), situé à moins de 10 kb du gène *KLHL25*, a été identifié comme significativement associé à la sévérité de l'affaissement de la paupière supérieure. Deux SNPs imputés, en fort déséquilibre de liaison avec rs4842897 ($r^2 > 0,9$), ont également été repérés.

Cette étude « génome entier » nous a donc permis de mettre en évidence des résultats très intéressants puisque l'implication des histones et du gène *KLHL25* dans le vieillissement global a déjà été avancée.

A Genome-Wide Association Study identifies new genetic associations in upper eyelid sagging.

Vincent Laville[1]*, Sigrid Le Clerc[1]*, Khaled Ezzedine[2,3], Randa Jdid[4], Lieng Taing[1], Toufik Labib[1], Cedric Coulonges[1], Damien Ulveling[1], Wassila Carpentier[5], Pilar Galan[2], Serge Hercberg[2,6], Frederique Morizot[4], Julie Latreille[4], Denis Malvy[2,7], Erwin Tschachler[8,9], Christiane Guinot[8,10], Jean-François Zagury[1]*[#].

[1] Équipe Génomique, Bioinformatique et Applications, Chaire de Bioinformatique, Conservatoire National des Arts et Métiers, Paris, France;
[2] UMR U557, INSERM/U1125 INRA/CNAM, University Paris 13/Centre de Recherche en Nutrition Humaine Ile-de-France, Bobigny, France;
[3] Department of Dermatology, Hôpital Saint-André, Bordeaux, France;
[4] Chanel R&T, Department of Skin Knowledge & Women Beauty, Pantin, France;
[5] Plate-forme Post-Génomique P3S, Hôpital Pitié-Salpêtrière, Paris, France;
[6] Department of Public Health, Hôpital Avicenne, Bobigny, France;
[7] Department of Internal Medicine and Tropical Diseases, Hôpital Saint-André, Bordeaux, France;
[8] CE.R.I.E.S.[§], Neuilly sur Seine, France;
[9] Department of Dermatology, University of Vienna Medical School, Vienna, Austria;
[10] Computer Science Laboratory, University François Rabelais of Tours, Tours, France;

[§] CE.R.I.E.S. is a research centre on human skin founded by CHANEL.
*: These authors share an equal contribution to the work.

[#]: Correspondance to:

Jean-François Zagury,
292 rue Saint Martin, 75003 Paris, France,
Tel: 33 1 58 80 88 20, e-mail: zagury@cnam.fr

Short Tittle: Genetics associations with upper eyelid sagging severity.

Abbreviations: GWAS, Genome Wide Association Study; LD, Linkage Disequilibrium; MAF, Minor Allele Frequency; SNP, Single Nucleotide Polymorphism.

Abstract

A genome-wide association study was conducted on a cohort of 502 French middle-aged women to identify genetic factors which may impact upper eyelid sagging. After Single Nucleotide Polymorphisms (SNP) quality controls, 795,063 SNPs remained for analysis purposes. The relationships between genotypes and upper eyelid sagging severity were measured separately using linear regressions adjusted for potential confounding factors. Imputation of the 1000 Genomes SNPs/indels was also performed.

Two signals were found associated with upper eyelid sagging severity on chromosome 10 and on chromosome 15. On chromosome 10, two genotyped SNPs, rs4746957 and rs16927253, passed the genome-wide significance threshold (5×10^{-8}). These polymorphisms, in high linkage disequilibrium, are located in an intronic part of the *H2AFY2* gene which encodes for a member of the H2A histone family. Moreover, SNPs in this region with low p-values were identified very close to the type XIII collagen *COL13A1* gene. On chromosome 15, one genotyped SNP, rs4842897, exhibited a p-value below the significance threshold. This SNP is located less than 10 kilobases away from the *KLHL25* gene.

These genes may constitute good candidate genes for the investigation of molecular mechanisms underlying upper eyelid sagging.

Introduction

Upper eyelid sagging, also called dermatochalasis or blepharoptosis, is a problematic consequence of skin aging characterized by a relaxation of skin and a large amount of abnormal elastic fibres (Lee *et al.*, 2012). In addition to its aesthetic concern for patients dermatochalasis can also be responsible for visual disorders and/or ocular irritations leading to a loss of visual field (Fay *et al.*, 2003). As for global skin ageing, eyelid drooping is the consequence of extrinsic and intrinsic factors (Jacobs *et al.*, 2014). Environmental factors involved in skin ageing, such as sun exposure, body mass index or smoking are well characterized. However, still little is known about intrinsic factors responsible for skin aging and eyelid drooping although the heritability percentage is estimated to 61% (Jacobs *et al.*, 2014). These past few years, several studies investigated the genetic mechanisms involved in skin ageing and highlighted the role of several genes such as *STXBP5L* (Le Clerc *et al.*, 2013) or *DIAPH2* (Chang *et al.*, 2014). Moreover, the association of eyelid drooping with several genetic disorders, such as cutis laxa (Berk *et al.*, 2012), Elhers Danlos syndrome (Packer and Blades, 1954) or amyloidosis (Gonnering and Sonneland, 1987), suggest an important role for genetic factors. A recent study dealing with both intrinsic and extrinsic factors associated with eyelid drooping in two independent Dutch and British cohorts was recently published (Jacobs *et al.*, 2014). This genome wide association study pointed to variants near the *TGFB1* gene in dermatochalasis severity.

In the present work, we described a genome wide association study performed on a unique cohort of 502 unrelated French women to look for genetic associations with eyelid drooping.

Results

Genotype data were obtained with the Illumina HumanOmni1-Quad BeadChips on a sample of 520 French middle-aged women from the SU.VI.MAX cohort. After quality control of the genotype data, a set of 795,063 genotyped SNPs was retained in 502 women. For each woman, a grade (ranging from 0 to 5) depicting upper eyelid sagging was defined from picture analysis. This final population of 502 women is described according to the severity of eyelid drooping (Table 1). As expected, age was a major marker for eyelid severity and we also computed the Kendall correlation coefficient which was $r = 0.39$ ($P < 0.0001$). For each SNP, we then looked for genetic associations by a linear regression adjusted for non-genetic skin aging factors (such as age and BMI) and for population stratification as well. Two significant signals ($p < 5.10^{-8}$) were found in chromosome 10 and chromosome 15 in the genotypic mode (Figure 1).

In chromosome 10, rs4746957 ($P = 1.69 \times 10^{-8}$) and rs16927253 ($P = 2.23 \times 10^{-10}$) were below the genome-wide significance threshold. According to the 1000 Genomes European sample, rs4746957 and rs16927253 are in total linkage disequilibrium (LD) and their minor allele frequencies (MAF) are about 6%. In our study, the two SNPs did not exhibit the same p-values because of different proportions of missing data. The genotype distribution in respect to the different grades suggested that the effect of the minor allele C was likely dominant and that it tends to prevent eyelid drooping (Figure 2). This observation was confirmed by the association p-values of rs4746957 and rs16927253 in the dominant model ($P_{DOM} = 9.08 \times 10^{-9}$ and $P_{DOM} = 1.29 \times 10^{-10}$). These two SNPs were located in an intronic part of the *H2AFY2* gene. After imputation, ten additional SNPs and one indel passed the significance threshold (Figure 1) and were in high LD with rs4746957 and rs16927253 ($r^2 > 0.8$). All these imputed SNPs were also located in intronic parts of the *H2AFY2* gene. A zoom on this region showed

that these significant SNPs are located near the *AIFM2*, *TYSND1*, and *SAR1A* genes. Moreover, other neighboring SNPs close to the *COL13A1* gene also exhibited low p-values (Figure 3).

In chromosome 15, rs4842897 exhibited a p-value ($P = 9.13 \times 10^{-12}$) below the significance threshold. The genotype distribution suggested that the minor allele A was associated with a less severe sagging and that the effect was likely recessive (Figure 4), as confirmed by the association p-value in the recessive mode ($P_{REC} = 7.08 \times 10^{-8}$). This SNP, with a minor allele frequency (MAF) of 15%, was located 8.4 kb away from the nearest gene *KLHL25*. Moreover, two imputed SNPs in high LD with rs4842897 ($r^2 > 0.9$) also exhibited significant p values. A zoom in the region shows isolated peaks (Figure 5), however the relatively high MAF and quality of imputation suggests that the signal observed should be real.

We looked in several eQTL, miRNA, and transcription databases, but none of the SNPs identified in this study was associated with an impact on gene expression.

Discussion

We have found several SNPs both on chromosome 10 and chromosome 15 significantly associated with upper eyelid severity. On chromosome 10, significant genotyped SNPs were located into intronic parts of the *H2AFY2* gene. This gene encodes a member of the H2A histone family called macroH2A2, which similarly to other histones, plays an important role in epigenetic organization and represses the transcription of chromatin domains in a subset of nucleosomes (Costanzi and Pehrson, 2001). This gene is expressed in the skin and is also associated with the skin disease lupus erythematosus. Several studies have already highlighted the role played by histones and epigenetics in cellular senescence and aging (Das and Tyler, 2013; Oberdoerffer, 2010; Rai and Adams, 2012). The *H2AFY2* gene exhibits differential expression patterns between aging and young cells and the macroH2A isoforms are highly expressed in senescent cells (Sporn *et al.*, 2009). This protein is also contained in senescence-associated heterochromatin foci (SAHF) which repress genes promoting proliferation and leads to an irreversible cell cycle exit (Zhang *et al.*, 2007). Overall, H2A2 has been little studied but appears as an important candidate gene for confirmatory studies. Three other genes were close to the significant SNPs. The *AIFM2* gene encodes for the apoptosis-inducing factor 2 whose expression induces apoptosis (Gong *et al.*, 2007). The *TYSND1* gene encodes a protease involved in the beta-oxidation of fatty acid (Kurochkin *et al.*, 2007). The *SAR1A* gene is involved in vesicle trafficking (Long *et al.*, 2010). Yet, we have not found evidence for particular relationship between these three genes and aging processes.

The last neighboring gene, *COL13A1*, encodes the alpha chain of the type XIII collagen, localized in the plasma membrane and that plays a role in the maintenance of connective tissues. *COL13A1* gene is also expressed in skin and fibroblasts, in the human eye and especially in the skeletal ciliary muscle and several nerves (Sandberg-Lall *et al.*, 2000).

Interestingly, a study on mice showed that type XIII collagen is involved in the adhesion between muscle fibre and basement membrane (Kvist *et al.*, 2001). Another study on mice demonstrated that this collagen is involved in the maturation of the skeletal neuromuscular function and showed a compromised function of the neuromuscular synapse in the absence of collagen XIII (Latvanlehto *et al.*, 2010). From these observations, *COL13A1* also appears as a gene of interest and two hypotheses could be drawn. First, a deficiency of the protein or a change in its expression level could lead to an abnormal level of fat in the connective tissue of the eyelid which could explain the sagging. Second, a modification in the neuromuscular junction in the eyelid could be responsible for the deficiency of the eyelid muscle leading to the sagging.

The chromosome 15 SNPs were located less than 10kb away from the *KLHL25* gene, 33kb away from the microRNA *MIR-1276* and 54 kb away from the *AKAP13* gene. The *KLHL25* gene encodes for the ectoderm-neural cortex protein 2. This gene is also expressed in connective tissues and especially in the adipose tissue. Very little is known about this protein and more generally on the Kelch-like gene family (Dhanoa *et al.*, 2013). Yet, it has already been associated with the biological age measured by an osseographic scoring system (Lunetta *et al.*, 2007). The *AKAP13* gene is a member of the kinase anchor proteins which binds to the regulatory subunit of a protein kinase. More specifically, this gene enhances ligand dependent activity of oestrogen receptors (Driggers *et al.*, 2001). Interestingly, several studies highlighted that oestrogen has a protective effect on skin aging, as this hormone prevents from a decrease in skin collagen after the menopause (Shah and Maibach, 2001). No relevant biological information has been found for the predicted targets of the *MIR-1276* microRNA.

The recent work by Jacobs *et al*. clearly demonstrated an important genetic component for eyelid drooping (Jacobs *et al.*, 2014). They also performed a GWAS and identified a recessive effect of the minor allele (C) of the rs11876749 SNP. We could not replicate this result in our

analysis. Nevertheless, it would be interesting to pool the data of the two studies and perform a meta-analysis.

In this work, we have identified several new genetic associations with eyelid sagging and found relevant biological hypotheses. As for any genetic association study, these results will have to be confirmed by replication in other cohorts or by direct biological experimentations.

Materials and methods

Study design and population

This cohort was more precisely described in a previous study (Le Clerc *et al.*, 2013). In the autumn/winter of 2002–2003, 570 of the 2,257 middle-aged women living in the Paris area included in the SU.VI.MAX cohort (Hercberg *et al.*, 1998) agreed to participate in a research on skin aging and provided informed consent. Each participant completed a self-administered questionnaire related to lifetime sun exposure behaviour and three standardized, high-resolution digital images (2,008 x 3,032 pixels) of the face were taken under normalized lightning conditions (one frontal view of the face and one of each profile), using a Kodak DCS 760 digital camera with a 105 mm camera lens (Kodak, Paris, France).

Outcome variables: phenotype analyzed

After image acquisition, upper eyelid drooping was evaluated by dermatologists using a specific ordinal grade, ranging from 0 to 5, with photographic illustrations (Morizot F, 2002). In this study, we focused on associations between SNPs genotypes and this particular grade.

Covariates used for the statistical analysis

To focus more specifically on the genetic factors affecting upper eyelid sagging, several characteristics susceptible to affect the phenotype were taken into account: age (in years), body mass index (BMI; in kgm^{-2}), smoking habits (never, former, and current smoker), and hormonal status (nonmenopausal, menopausal with hormone replacement therapy, and menopausal without hormone replacement therapy). BMI was categorized as underweight or normal ($BMI < 25\ kgm^{-2}$), overweight ($25 \leq BMI < 30\ kgm^{-2}$), or obese ($BMI \geq 30\ kgm^{-2}$)

according to the World Health Organization (WHO) recommendations (WHO, 1995). Lifetime sun exposure intensity was estimated by a score based on data collected by a self-reported questionnaire. The design, validation, and description of this score have been described previously (Guinot *et al.*, 2001).

Genotyping method

The genotyping method for this cohort has already been precisely described (Le Clerc *et al.*, 2013). The 529 women were genotyped using Illumina Infinium HumanOmni1-Quad BeadChips (Illumina, San Diego, CA) that contain 1,140,419 markers and a sample of 250 ng of ADN by individual was used to obtain genotypes. For the analysis, we considered only SNPs, consequently excluding the copy-number variations that represented 91,706 markers on the HumanOmni1-Quad BeadChips. Moreover, 2,182 SNPs were removed because they were located on the Y chromosome and they could not be analyzed as the population was composed of women.

Quality control

Using the GenomeStudio software (version 1.6.3; Illumina), we analyzed the crude genotyping data, and SNPs were filtered according to several criteria. First, nine samples with a call rate (percentage of SNPs genotyped by sample) of <95% in the Illumina clusters were removed. Second, the SNPs with a call frequency (percentage of samples genotyped by SNP) of < 99% were reclustered. Third, after reclustering, samples with a call rate of <98% were deleted. This method has been already used in several studies (Le Clerc *et al.*, 2009; Limou *et al.*, 2010; Limou *et al.*, 2009). The clustering step can create SNP genotyping errors, which can be prevented by following the Illumina procedure

(http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf). This method evaluates the quality of the newly created clusters according to several criteria, which can be manually checked and corrected as necessary. In total, after all the quality control steps were carried out, 56,479 SNPs with a call frequency of < 98% (2% of missing data) were excluded. This procedure ensures reliable genotyping data with little missing data. Hardy–Weinberg equilibrium analysis was performed for each SNP in each group by using an exact statistical test implemented in the PLINK software (Purcell *et al.*, 2007). Deviation from Hardy–Weinberg equilibrium in a group of patients suggests an error in genotyping. Thus, 3,866 SNPs, which were not in the Hardy–Weinberg equilibrium ($P < 5 \times 10^{-3}$), were rejected in this way. We removed 191,123 SNPs with minor allele frequency < 1% to avoid error of genotyping, leaving a total of 795,063 SNPs.

Identification of population stratification

To correct for possible population stratification, genotypes were analyzed using the EIGENSTRAT utility of the EIGENSOFT package version 4.2 (Price *et al.*, 2006). The two first pass with the Eigenstrat software pointed out 18 outliers, who were removed from further analyses. Then, a third pass without outliers was performed to determine the Eigen vectors. In the statistical analysis, we used the top two Eigen vectors as covariates to correct for population substructure in the association analyses.

Statistical analysis

Of the 570 women who participated in the study, 68 were excluded from the analysis: 18 had a history of recent antiaging invasive procedures and 10 were observably non-Caucasian. In addition, one sample was removed because of insufficient DNA concentration, 12 samples

were removed because the DNA was damaged, and nine samples were removed after quality control. Furthermore, 18 outliers appeared during the stratification analysis. Thus, the population investigated for our genome-wide association study was composed of 502 individuals. In addition, Kendall rank correlation coefficient was calculated between age and the outcome variable. Then, for each of the remaining 795,063 SNPs, the associations between the genotypes and upper eyelid sagging were measured by a multivariate linear regression adjusted on age, smoking habits, BMI, hormonal status, lifetime sun exposure intensity and the two first Eigenstrat principal components and the potential confounding factors. The genome-wide significance threshold (5×10^{-8}) was used to assess statistical significance. Finally, the imputed data were similarly analyzed using the SNPTEST software (Marchini *et al.*, 2007). Only SNPs with imputation quality score > 0.8, MAF > 1% and test score > 0.9 were kept.

Imputation for high density mapping

The genotype data were first phased using Shape-IT2 (Delaneau *et al.*, 2013). The phased data were then imputed using IMPUTE2 (Marchini *et al.*, 2007). As reference haplotypes, we used genotype data of 1094 individuals from the phase I integrated variant set of the 1000 Genomes project released in March 2012 and updated in August 2012 (Abecasis *et al.*, 2010). To reduce uncertainty, SNPs with MAF < 1% in European samples of the reference panel were not used to impute our data.

Bioinformatics exploration

To further explore the signals observed by the GWAS, we tried to look for modifications in mRNA expression levels (Genevar (Nica *et al.*, 2011; Yang *et al.*, 2010), Dixon (Dixon *et al.*,

2007) databases and GHS-Express (Zeller *et al.*, 2010) databases), splicing (NetGene2, http://www.cbs.dtu.dk/services/NetGene2/), polyadenylation regions (polyAH, http://linux1.softberry.com/berry.phtml?topic=polyah&group =programs&subgroup=promoter and polyApred, http://www.imtech. res.in/raghava/polyapred/submission.html), transcription factor binding sites (SignalScan, http://www-bimas.cit.nih.gov/molbio/signal/, TESS, http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=WELCOME, and TF Search, http://www.cbrc.jp/research/db/TFSEARCH.html, derived from TRANSFAC database), and miRNA genes or miRNA targets (miRBAse, http://www.mirbase.org/, miRTarBase, http://mirtarbase. mbc.nctu.edu.tw/, MicroCosm Targets, http://www.ebi.ac.uk/enrightsrv/microcosm/htdocs/targets/v5/)

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

The authors gratefully acknowledge the dedicated efforts of all the SU.VI.MAX volunteers, the investigators, and the staff members involved in this study, especially Dr Sandrine Bertrais, and Ms Nathalie Arnault and Mr Gwenael Monot who coordinated the data management.

Bibliography

Abecasis GR, Altshuler D, Auton A*, et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061-73.

Berk DR, Bentley DD, Bayliss SJ*, et al.* (2012) Cutis laxa: a review. *J Am Acad Dermatol* 66:842 e1-17.

Chang AL, Atzmon G, Bergman A*, et al.* (2014) Identification of genes promoting skin youthfulness by genome-wide association study. *J Invest Dermatol* 134:651-7.

Costanzi C, Pehrson JR (2001) MACROH2A2, a new member of the MARCOH2A core histone family. *J Biol Chem* 276:21776-84.

Das C, Tyler JK (2013) Histone exchange and histone modifications during transcription and aging. *Biochim Biophys Acta* 1819:332-42.

Delaneau O, Zagury JF, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10:5-6.

Dhanoa BS, Cogliati T, Satish AG*, et al.* (2013) Update on the Kelch-like (KLHL) gene family. *Hum Genomics* 7:13.

Dixon AL, Liang L, Moffatt MF*, et al.* (2007) A genome-wide association study of global gene expression. *Nat Genet* 39:1202-7.

Driggers PH, Segars JH, Rubino DM (2001) The proto-oncoprotein Brx activates estrogen receptor beta by a p38 mitogen-activated protein kinase pathway. *J Biol Chem* 276:46792-7.

Fay A, Lee LC, Pasquale LR (2003) Dermatochalasis causing apparent bitemporal hemianopsia. *Ophthal Plast Reconstr Surg* 19:151-3.

Gong M, Hay S, Marshall KR*, et al.* (2007) DNA binding suppresses human AIF-M2 activity and provides a connection between redox chemistry, reactive oxygen species, and apoptosis. *J Biol Chem* 282:30331-40.

Gonnering RS, Sonneland PR (1987) Ptosis and dermatochalasis as presenting signs in a case of occult primary systemic amyloidosis (AL). *Ophthalmic Surg* 18:495-7.

Guinot C, Malvy D, Latreille J *et al*. (2001). Sun exposure behaviour of a general adult population in France. In: Ring J, Weidinger S, Darsow U eds *Skin and Environment - Perception and Protection*. Monduzzi editore S.p.A: Bologna, 1099–106

Hercberg S, Galan P, Preziosi P*, et al.* (1998) Background and rationale behind the SU.VI.MAX Study, a prevention trial using nutritional doses of a combination of antioxidant vitamins and minerals to reduce cardiovascular diseases and cancers. SUpplementation en VItamines et Mineraux AntioXydants Study. *Int J Vitam Nutr Res* 68:3-20.

Jacobs LC, Liu F, Bleyen I*, et al.* (2014) Intrinsic and Extrinsic Risk Factors for Sagging Eyelids. *JAMA Dermatol*.

Kurochkin IV, Mizuno Y, Konagaya A*, et al.* (2007) Novel peroxisomal protease Tysnd1 processes PTS1- and PTS2-containing enzymes involved in beta-oxidation of fatty acids. *EMBO J* 26:835-45.

Kvist AP, Latvanlehto A, Sund M*, et al.* (2001) Lack of cytosolic and transmembrane domains of type XIII collagen results in progressive myopathy. *Am J Pathol* 159:1581-92.

Latvanlehto A, Fox MA, Sormunen R*, et al.* (2010) Muscle-derived collagen XIII regulates maturation of the skeletal neuromuscular junction. *J Neurosci* 30:12230-41.

Le Clerc S, Limou S, Coulonges C*, et al.* (2009) Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J Infect Dis* 200:1194-201.

Le Clerc S, Taing L, Ezzedine K*, et al.* (2013) A genome-wide association study in Caucasian women points out a putative role of the STXBP5L gene in facial photoaging. *J Invest Dermatol* 133:929-35.

Lee H, Park M, Lee J*, et al.* (2012) Histopathologic findings of the orbicularis oculi in upper eyelid aging: total or minimal excision of orbicularis oculi in upper blepharoplasty. *Arch Facial Plast Surg* 14:253-7.

Limou S, Coulonges C, Herbeck JT*, et al.* (2010) Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS. *J Infect Dis* 202:908-15.

Limou S, Le Clerc S, Coulonges C*, et al.* (2009) Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J Infect Dis* 199:419-26.

Long KR, Yamamoto Y, Baker AL*, et al.* (2010) Sar1 assembly regulates membrane constriction and ER export. *J Cell Biol* 190:115-28.

Lunetta KL, D'Agostino RB, Sr., Karasik D*, et al.* (2007) Genetic correlates of longevity and selected age-related phenotypes: a genome-wide association study in the Framingham Study. *BMC Med Genet* 8 Suppl 1:S13.

Marchini J, Howie B, Myers S*, et al.* (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39:906-13.

Morizot F LS, Guinot C, Binder M, Tschachler E (2002) Development of photographic scales documenting features of skin ageing based on digital images. *Ann Dermatol Venereol*.

Nica AC, Parts L, Glass D*, et al.* (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* 7:e1002003.

Oberdoerffer P (2010) An age of fewer histones. *Nat Cell Biol* 12:1029-31.

Packer BD, Blades JF (1954) Dermatorrhexis: a case report; the so-called Ehlers-Danlos syndrome. *Va Med Mon (1918)* 81:27-30.

Price AL, Patterson NJ, Plenge RM*, et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904-9.

Purcell S, Neale B, Todd-Brown K*, et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559-75.

Rai TS, Adams PD (2012) Lessons from senescence: Chromatin maintenance in non-proliferating cells. *Biochim Biophys Acta* 1819:322-31.

Sandberg-Lall M, Hagg PO, Wahlstrom I*, et al.* (2000) Type XIII collagen is widely expressed in the adult and developing human eye and accentuated in the ciliary muscle, the optic nerve and the neural retina. *Exp Eye Res* 70:401-10.

Shah MG, Maibach HI (2001) Estrogen and skin. An overview. *Am J Clin Dermatol* 2:143-50.

Sporn JC, Kustatscher G, Hothorn T*, et al.* (2009) Histone macroH2A isoforms predict the risk of lung cancer recurrence. *Oncogene* 28:3423-8.

Yang TP, Beazley C, Montgomery SB*, et al.* (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics* 26:2474-6.

Zeller T, Wild P, Szymczak S*, et al.* (2010) Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One* 5:e10693.

Zhang R, Chen W, Adams PD (2007) Molecular dissection of formation of senescence-associated heterochromatin foci. *Mol Cell Biol* 27:2343-58.

Tables

Table 1. Description of the population according to the eyelid drooping severity

	Severity of upper eyelid drooping					
	Grade 0/1[1] N=29	Grade 2 N=107	Grade 3 N=192	Grade 4/5[1] N=174	Total N=502	P value of test
Age (years)	51.0±5.0[2]	54.0±6.	57.3±5.6	61.2±5.3	57.6±6.4	< 0.0001[4]
Lifetime sun exposure (score)	5.1±3.3	5.5±3.6	5.0±3.6	5.6±3.5	5.3±3.5	0.36[4]
BMI classification						0.20[5]
Normal	20 (4.0)[3]	82	126 (25.1)	107 (21.3)	335 (66.7)	
Overweight	7 (1.4)	20 (4.0)	45 (9.0)	50 (10.0)	122 (24.3)	
Obese	2 (0.3)	5 (1.0)	21 (4.2)	17 (3.4)	45 (9.0)	
Hormonal status						< 0.0001[5]
No menopausal	15 (3.0)	39 (7.8)	35 (7.0)	5 (0.6)	94 (18.7)	
Menopausal with HRT	10 (2.0)	45(9.0)	110 (22.0)	97 (19.3)	262 (52.2)	
Menopausal without HRT	4 (0.8)	23 (4.6)	47 (9.4)	72 (14.3)	146 (29.1)	
Smoking habits						0.18[5]
Never	16 (3.2)	62	104 (20.7)	104 (20.7)	286 (57.1)	
Former smoker	8 (1.6)	32 (6.4)	60 (12.0)	60 (12.0)	160 (31.9)	
Current smoker	5 (1.0)	13 (2.6)	28 (1.6)	10 (2.0)	55 (11.0)	
Eye color						0.25[5]
Blue/grey	9 (1.8)	21 (4.2)	54 (10.8)	52 (10.4)	136 (27.3)	
Green/hazel/brown/black	20(4.0)	86	135 (27.1)	122 (24.4)	363 (72.7)	
Hair color at twenty						0.02[5]
Blond/red	4 (0.8)	21 (4.2)	32 (6.4)	51 (10.2)	108 (21.6)	
Light&dark brown/black	25 (5.0)	86	157 (31.5)	123(24.6)	391 (78.4)	
Skin color without tanning						0.05[5]
Fair	23 (4.6)	77	141 (28.3)	147 (29.5)	388 (77.8)	
Dark	6 (1.2)	30 (6.0)	48 (9.6)	27 (5.4)	111 (22.2)	
History of facial freckles						< 0.39[5]
No	18 (3.6)	70	110 (22.0)	96 (19.2)	294 (58.9)	
Yes	11 (2.2)	37 (7.4)	79 (15.8)	78 (15.6)	205 (41.1)	
Suntan intensity						0.25[5]
None/slight/light	17 (3.4)	56	115 (23.0)	112 (22.4)	300 (60.1)	
Dark/very dark	12 (2.4)	51	74 (14.8)	62 (12.4)	199 (39.9)	
Sunburn event frequency						0.57[5]
None/rare	19 (3.8)	73	141 (28.3)	125 (25.1)	358 (71.7)	
Frequent/constant	10 (2.0)	34 (6.8)	48 (9.6)	49 (9.8)	141 (28.3)	

[1] As too few people had a grade 1 and 5, groups 0 and 1 and groups 4 and 5were gathered.
[2] Mean ± Standard Deviation
[3] Frequency and (%), due to possible missing values the sum of the cell frequencies can be smaller than the total indicated in the top of the columns.
[4] ANOVA test
[5] χ^2 test
Abbreviations: BMI, Body Mass Index, HRT, Hormone Replacement Therapy

Figures

Figure 1. Manhattan plot of the association study with the upper eyelid sagging grade

Distribution of $-\log_{10}(P)$ obtained for the associations tested between genotypes and upper eyelid sagging along the human chromosome.

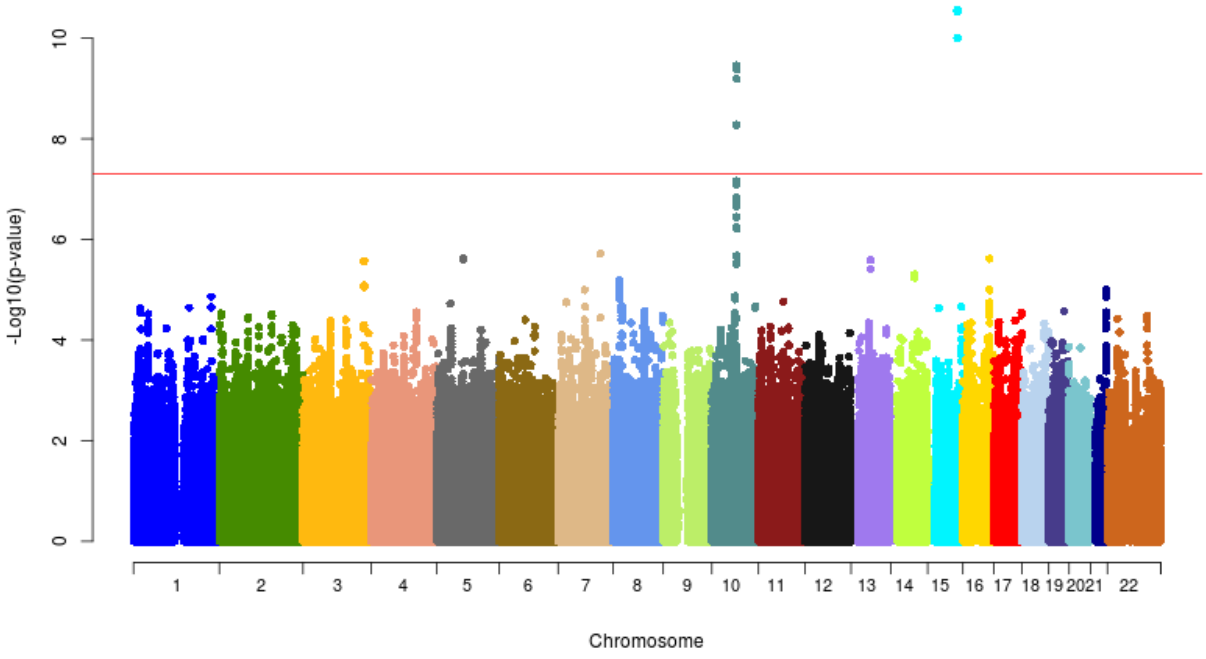


Figure 2. Genotype distribution of rs16927253 in respect to the upper eyelid sagging severity

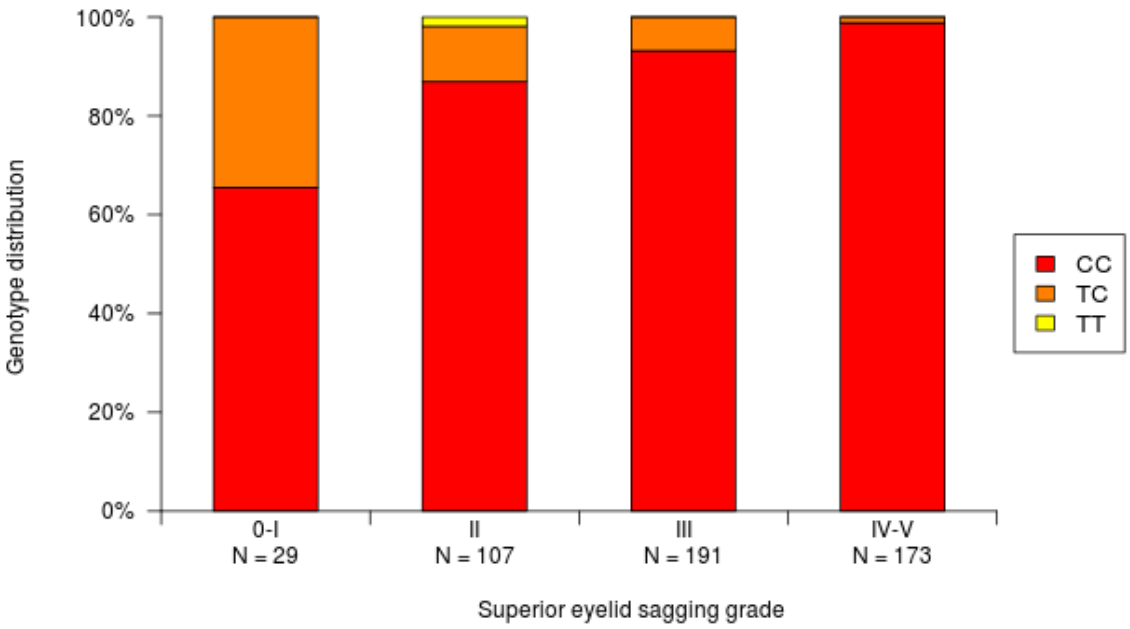


Figure 3. Genetic context of the SNPs identified on the chromosome 10

Distribution of the $-\log_{10}(P)$ obtained for the associations tested between genotypes and upper eyelid sagging in the chromosome 10 region located between 71.55 Mb and 71.90 Mb. The linkage disequilibrium values between the top SNP rs16927253 (purple) and other SNPs are represented by colored points according to the LD legend in the left top. A genetic map of the region is represented below the distribution.

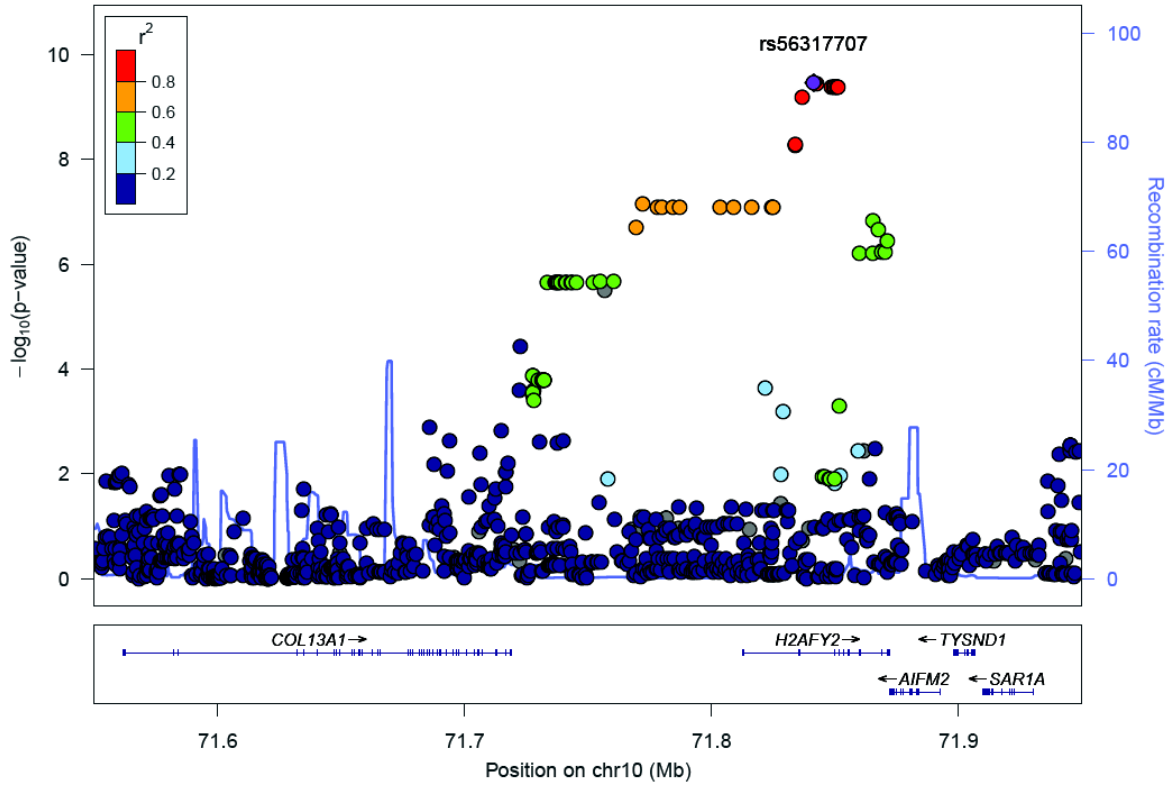


Figure 4. Genotype distribution of rs4842897 in respect to the upper eyelid sagging severity

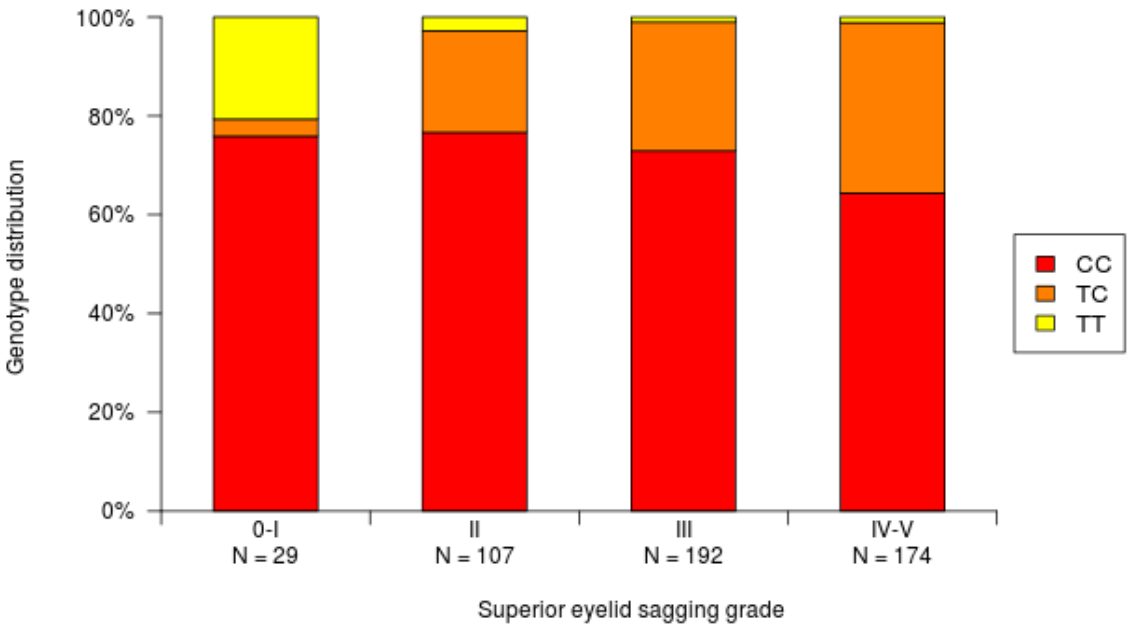
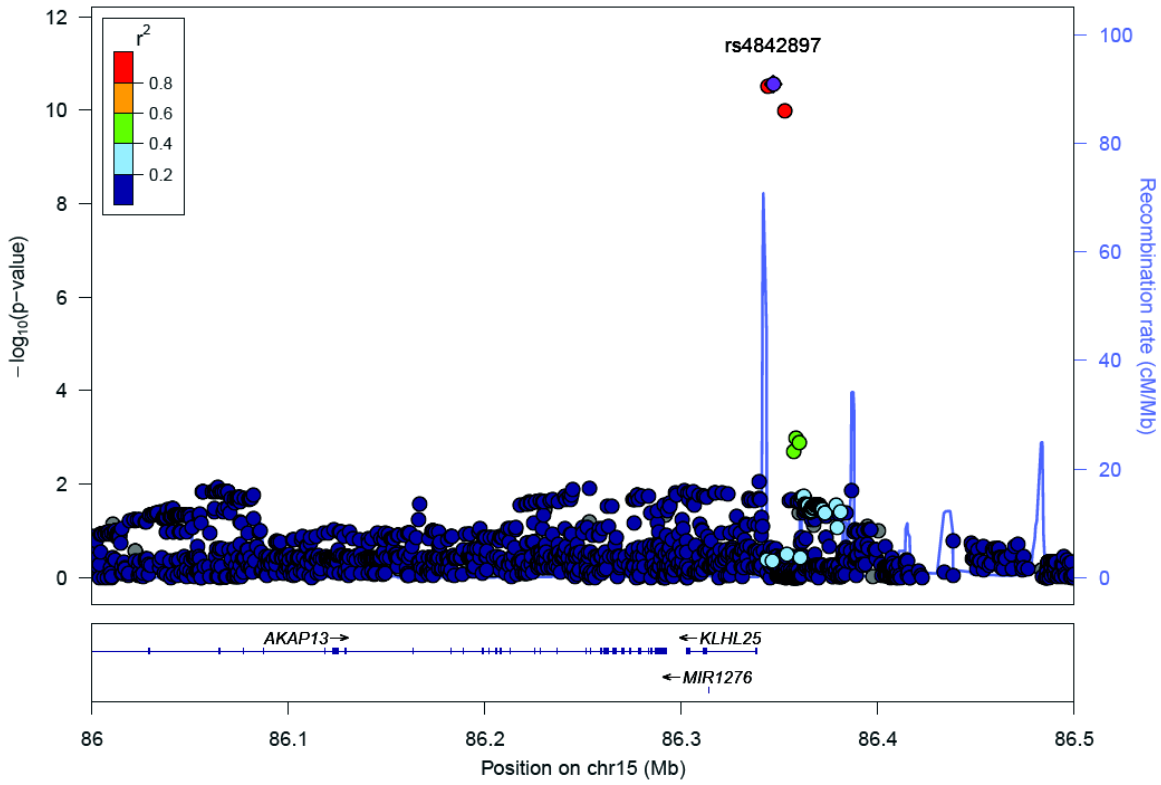


Figure 5. Genetic context of the SNPs identified on the chromosome 15

Distribution of the $-\log_{10}(P)$ obtained for the associations tested between genotypes and upper eyelid sagging in the chromosome 15 region located between 86 Mb and 86.5 Mb. The linkage disequilibrium values between the top SNP rs4842897 (purple) and other SNPs are represented by colored points according to the LD legend in the left top. A genetic map of the region is represented below the distribution.



3 Recherche de voies de signalisation biologiques significativement associées avec des indicateurs de vieillissement cutané

3.1 Contexte

Les résultats précédents soulignent l'efficacité des analyses « génome entier » pour mettre en évidence des signaux individuels forts, mais elles restent inefficaces pour mettre en évidence des effets à petite échelle de variants génétiques (PMID : 21085203). Nous avons mis en œuvre une autre approche, actuellement en plein développement, qui consiste à utiliser l'information connue sur les voies de signalisation et les réseaux biologiques. Cette méthode permet d'inclure des connaissances biologiques a priori pour essayer d'identifier des fonctions biologiques globales potentiellement impliquées dans le vieillissement cutané, en tenant compte des variants ayant des effets plus faibles.

Nous avons donc entrepris l'identification de voies de signalisation pouvant impacter 4 indicateurs importants du vieillissement cutané sur le visage à notre disposition : un grade de photo-vieillissement cutané, un score de rides sur le visage, un score de lentigines sur le visage, et un score mesurant le relâchement de la peau sur le visage.

Il existe différentes méthodes pour sélectionner des voies de signalisation/réseaux biologiques capables d'influencer un phénotype. Nous avons choisi d'appliquer l'algorithme Gene Set Enrichment Analysis (GSEA) [183] qui vise à identifier les voies de signalisation enrichies en faibles p-valeurs comparativement aux autres voies de signalisation (voir Matériel et Méthodes, chapitre 2.4). Les voies de signalisations que nous avons étudiées sont issues de la base de données KEGG et sont constituées de 20 à 200 gènes.

3.2 Résultats

Les associations ont été recherchées entre les voies de signalisation et chacun des quatre indicateurs du vieillissement cutané sur le visage. Des associations significatives ont été obtenues avec 18 voies de signalisation pour le grade de Larnier de photo-vieillissement global (FDR < 25%) (Tableau 9).

Nom du pathway	Nombre de gènes	FDR	
Nucleotide excision repair	37	0.012	Genetic Information Processing
Selenoamino acid metabolism	30	0.042	Metabolism
Thyroid cancer	31	0.043	Disease
Wnt signaling pathway	144	0.048	Environmental Information Processing
Primary immunodeficiency	31	0.055	Disease
Taste transduction	44	0.106	Organismal Systems
Regulation of actin cytoskeleton	185	0.114	Cellular Processes
Melanogenesis	102	0.121	Organismal Systems
Inositol phosphate metabolism	48	0.125	Metabolism
Base excision repair	30	0.138	Genetic Information Processing
Systemic lupus erythematosus	50	0.144	Disease
Long-term potentiation	64	0.184	Organismal Systems
Proteasome	21	0.199	Genetic Information Processing
Phosphatidylinositol signaling system	76	0.203	Environmental Information Processing
Mismatch repair	21	0.212	Genetic Information Processing
O-Glycan biosynthesis	28	0.214	Metabolism
Ubiquitin mediated proteolysis	119	0.224	Genetic Information Processing

Tableau 9. Liste des voies de signalisation significativement associées (FDR < 25%) au photo-vieillissement cutané. Les voies de signalisation situées au dessus de la ligne pointillée rouge ont un FDR inférieur à 5%. La couleur des lignes permet la distinction des différentes catégories de voies de signalisation selon la classification de KEGG.

Parmi ces associations, 4 voies de signalisation avaient un FDR < 5% : la réparation de l'ADN par excision de nucléotide (FDR = 1,2%), le métabolisme des acides aminés contenant du sélénium (FDR = 4,2%), le cancer de la thyroïde (FDR = 4,3%) et la voie de signalisation Wnt (FDR = 4,8%). Cette dernière voie de signalisation constitue un résultat pertinent en raison de son implication dans la mélanogénèse, également associée au photo-vieillissement (FDR = 12,1%). De plus, une étude transcriptomique avait déjà mis en évidence la différence d'expression de 4 gènes appartenant à cette voie de signalisation chez les personnes âgées [149]. Enfin, 5 voies de signalisation sont liées au traitement de l'information génétique, et en particulier à la réparation des lésions de l'ADN. Cette observation pourrait s'expliquer par l'influence majeure de l'exposition aux rayons UV dans le photo-vieillissement qui sont responsables de dommages au niveau de l'ADN.

Des associations significatives avec 16 voies de signalisation biologiques ont été identifiées pour le score global de relâchement cutané (FDR < 25%) (Tableau 10).

Nom du pathway	Nombre de gènes	FDR	
Ribosome	61	0.076	Genetic Information Processing
Nucleotide excision repair	37	0.078	Genetic Information Processing
mTOR signaling pathway	44	0.080	Environmental Information Processing
Insulin signaling pathway	126	0.142	Organismal Systems
O-Glycan biosynthesis	28	0.146	Metabolism
DNA replication	33	0.150	Genetic Information Processing
Cell cycle	100	0.150	Cellular processes
Melanogenesis	102	0.150	Organismal Systems
Small cell lung cancer	83	0.151	Disease
Glycerophospholipid metabolism	59	0.158	Metabolism
Adipocytokine signaling pathway	67	0.159	Organismal Systems
Regulation of actin cytoskeleton	185	0.161	Cellular Processes
Basal transcription factors	28	0.167	Genetic Information Processing
Glycerolipid metabolism	43	0.173	Metabolism
Biosynthesis of unsaturated fatty acids	22	0.196	Metabolism
PPAR signaling pathway	65	0.205	Organismal Systems
VEGF signaling pathway	72	0.222	Environmental Information Processing
Folate biosynthesis	36	0.224	Metabolism
Melanoma	66	0.234	Disease

Tableau 10. Liste des voies de signalisation significativement associées (FDR < 25%) au score global de relâchement cutané. La couleur des lignes permet la distinction des différentes catégories de voies de signalisation selon la classification de KEGG.

Aucune voie de signalisation n'avait un FDR < 5%. Les réseaux biologiques les plus fortement associés au relâchement cutané étaient le ribosome (FDR = 7,6%) et la réparation de l'ADN par excision d'une base (FDR = 7,8%). La voie de signalisation mTOR (FDR = 8,0%), enzyme régulant de nombreux processus cellulaires comme la croissance ou la prolifération, est associée au phénotype. Cette association est très pertinente puisque l'implication de cette voie dans le vieillissement a été démontrée [185]. Enfin, il est intéressant de remarquer les associations de voies de signalisation (adipocytokines, FDR = 15,9% ; PPAR, FDR = 20,5%) et de voies métaboliques (glycérophospholipides, FDR = 15,8% ; glycérolipides, FDR = 17,3%) liées aux lipides qui sont un constituant majeur de la peau.

Le score global de rides a pu être associé de façon significative avec 10 voies de signalisation (Tableau 11).

Nom du pathway	Nombre de gènes	FDR	
Ribosome	61	0.019	Genetic Information Processing
Nucleotide excision repair	37	0.025	Genetic Information Processing
Proteasome	21	0.030	Genetic Information Processing
Biosynthesis of steroids	21	0.033	Metabolism
Primary immunodeficiency	31	0.034	Disease
Bladder cancer	39	0.067	Disease
Inositol phosphate metabolism	48	0.125	Metabolism
Homologous recombination	26	0.159	Genetic Information Processing
Cell cycle	100	0.173	Cellular processes
Glycerophospholipid metabolism	59	0.174	Metabolism

Tableau 11. Liste des voies de signalisation significativement associées (FDR < 25%) au score global de rides sur le visage. Les voies de signalisation situées au dessus de la ligne pointillée rouge ont un FDR inférieur à 5%. La couleur des lignes permet la distinction des différentes catégories de voies de signalisation selon la classification de KEGG.

Parmi ces 10 voies de signalisation, 5 d'entre elles avaient un FDR < 5%. Les 3 réseaux biologiques les plus fortement associés au score de rides : le ribosome (FDR = 1,9%), la réparation de l'ADN par excision d'une base (FDR = 2,5%), le protéasome (FDR = 3,0%), la biosynthèse des stéroïdes (FDR = 3,3%) et l'immunodéficience primaire (FDR = 3,4%). Les trois premières appartiennent à la catégorie du traitement de l'information génétique. Enfin, plusieurs voies de signalisation liées aux lipides ou à leur métabolisme sont également significativement associées au phénotype.

Enfin, l'étude des associations avec le score global de lentigines a permis de mettre en évidence 17 résultats significatifs (FDR < 25%) (Tableau 12).

Nom du pathway	Nombre de gènes	FDR	
Long-term depression	73	0.067	Organismal Systems
Thyroid cancer	31	0.099	Disease
Regulation of actin cytoskeleton	185	0.187	Cellular Processes
Systemic lupus erythematosus	50	0.207	Disease
Long-term potentiation	64	0.215	Organismal Systems
Complement and coagulation cascades	67	0.217	Organismal Systems
Methionine metabolism	21	0.219	Metabolism
Mismatch repair	21	0.226	Genetic Information Processing
Melanogenesis	102	0.226	Organismal Systems
Alanine and aspartate metabolism	32	0.230	Metabolism
Phosphatidylinositol signaling system	76	0.231	Environmental Information Processing
Fc epsilon RI signaling pathway	75	0.231	Organismal Systems
Arginine and proline metabolism	32	0.232	Metabolism
GnRH signaling pathway	94	0.232	Organismal Systems
Glycan structures - biosynthesis 1	112	0.233	Metabolism
Calcium signaling pathway	162	0.244	Environmental Information Processing
Natural killer cell mediated cytotoxicity	113	0.244	Organismal Systems

Tableau 12. Liste des voies de signalisation significativement associées (FDR < 25%) au score de lentigines solaires sur le visage. La couleur des lignes permet la distinction des différentes catégories de voies de signalisation selon la classification de KEGG.

Aucune voie de signalisation n'avait un FDR < 5%. Les voies de signalisation significativement associées appartiennent en grande partie (41%) à la catégorie « Systèmes de l'organismes ». En particulier, le système immunitaire semble lié aux lentigines puisque 3 voies de signalisation associées aux lentigines font partie de ce système : cascades de complément et de coagulation (FDR = 21,7%), la voie de signalisation Fc epsilon RI (FDR = 23,1%) et la cytotoxicité médiée par les lymphocytes Natural Killer (FDR = 24,4%). La mélanogenèse est également associée aux lentigines (FDR = 22,6%). Enfin, la voie de signalisation de l'hormone de libération des gonadotrophines hypophysaires (GnRH), déjà impliquée dans le vieillissement général [186], est également associée aux lentigines (FDR = 23,2%).

Notre étude utilisant l'approche dite de « voies de signalisations biologiques » nous a donc permis d'identifier des résultats significatifs pour chacun des indicateurs de vieillissement cutané utilisés. Parmi ceux-ci, la voie de signalisation de la mélanogenèse apparaît comme un signal particulièrement puissant puisque l'association avec cette voie a été retrouvée pour 3 de nos 4 indicateurs de vieillissement cutané. Ce résultat attendu a priori atteste alors de la bonne qualité des données analysées. De même, différentes voies de signalisation biologiques

impliquées dans l'immunité se sont révélées être significativement associées avec tous les indicateurs de vieillissement cutané à l'exception du score global de rides. Ces résultats permettent donc de fournir de nouvelles pistes pour mieux comprendre les mécanismes du vieillissement cutané. Nous avons choisi la méthode GSEA pour réaliser notre étude puisqu'il s'agit d'une approche rapide et reconnue. Cependant, nous souhaitons reproduire cette étude avec d'autres méthodes, notamment avec la méthode « SNP Ratio Test » et d'autres plus récentes et plus puissantes pour s'assurer de la réplication des signaux identifiés, et éventuellement détecter de nouvelles voies de signalisation reliées au vieillissement cutané.

Quatrième partie
Discussion et perspectives

1 Bilan des études d'association réalisées sur le vieillissement cutané

1.1 Travaux portant sur les lentigines

1.1.1 Rappels des résultats

Nous avons mené une étude d'association « génome entier » à la recherche de gènes impliqués dans la sévérité des lentigines.

Lorsque nous nous sommes intéressés aux lentigines situées sur le front, deux associations se sont révélées significatives. Il s'agit des polymorphismes rs9350204 et rs9358294, tous deux situés dans une partie intergénique de la région 6p22 du chromosome 6, à proximité du pseudogène *RPL29P17* et du gène *MBOAT1*.

Aucun résultat significatif n'a pu être mis en évidence en étudiant la sévérité des lentigines sur les joues.

Des associations significatives ont été obtenues avec le score global des lentigines sur le visage. En effet, 9 SNPs, répartis en deux groupes distincts et indépendants, présentaient un FDR inférieur à 25%. Ces deux groupes sont localisés sur le chromosome 6 dans les régions 6p21 et 6p22. Les SNPs identifiés dans la région 6p22 étaient les 2 mêmes SNPs significativement associés avec le grade de lentigines sur le front. Les 7 polymorphismes identifiés dans la région 6p21 sont situés dans la région *HLA*. Les individus hétérozygotes pour ces SNPs présentaient une sévérité des lentigines plus importante. Ces SNPs jouent un rôle dans l'expression de plusieurs gènes de la région HLA et en particulier de *HLA-C*. Enfin, l'imputation des allèles du *HLA* de classe I et de classe II a permis d'identifier une tendance d'association entre l'allèle *HLAC*0701* et le phénotype.

De plus, nous avons établi, à partir de la bibliographie, une liste de 38 gènes candidats (de laquelle le gène *MC1R* a été exclus car précédemment étudié) précédemment associés aux lentigines solaires. A partir des p-valeurs obtenues dans la GWAS, 19 SNPs sur les 851 SNPs provenant de ces 38 gènes présentaient un FDR inférieur à 25% :

- 12 sont situés dans le gène *MITF*, codant pour un facteur de transcription pour la mélanogenèse ;
- 2 dans le gène *OCA2* ;
- 1 dans les gènes *SCEL*, *AMBP* et *RBBP6*.

1.1.2 Comparaison avec la littérature

L'étude d'association « génome-entier » avec les lentigines que nous avons réalisée est la première et aucune étude génome-entier n'avait précédemment été publiée sur les lentigines à notre connaissance.

Plusieurs études d'association sur gènes-candidats ou de transcriptome sur cellules issues de lentigines ont été réalisées et avaient identifié des gènes potentiellement impliqués dans les lentigines solaires (voir Introduction, paragraphe 6.3.3). Ce sont d'ailleurs ces gènes que nous ont utilisés comme gène-candidats dans le cadre de notre GWAS.

L'association du gène *MC1R* avec les lentigines solaires constitue un résultat bien connu [158]. L'association entre ce gène et la sévérité des lentigines sur le visage avait déjà été étudiée dans notre cohorte lors d'une étude précédente [184] et aucune association significative n'avait été établie.

1.1.3 Interprétation biologique

La protéine HLA-C appartient aux molécules du HLA de classe I et joue un rôle essentiel dans l'immunité en présentant les antigènes aux lymphocytes cytotoxiques et en inhibant les cellules Natural Killer. Ces résultats, et plus particulièrement la sous-expression du *HLA-C* induite par certains variants génétiques identifiés, suggèrent une implication du système immunitaire dans la sévérité des lentigines. Au contraire, une forte expression de ce gène a été associée à un meilleur contrôle des carcinomes induits par les cellules squameuses. Ainsi, ces observations pourraient suggérer un mécanisme de reconnaissance des cellules cutanées anormales via le *HLA-C*, et les mélanocytes aberrants ne seraient alors pas éliminés par les cellules immunitaires lorsque ce gène est sous-exprimé.

En utilisant les gènes-candidats connus pour leur implication dans les lentigines solaires, nous avons pu retrouver des signaux d'intérêt (FDR<25%) pour les gènes suivants : *MITF*, *OCA2*, *SCEL, AMBP* et *RBBP6*. Le gène *MITF* est pertinent car très connu pour son rôle clé dans la mélanogenèse [187]. Le gène *OCA2* joue lui aussi un rôle important dans la pigmentation car

il serait impliqué dans la biogénèse de la mélanine [188]. Le gène *SCEL* code pour la scielline qui est un constituant de l'enveloppe cornée de kératinocytes [189]. Le gène *AMBP* code pour des protéines impliquées dans l'immunité et présentes dans la matrice extra-cellulaire [190]. Enfin, le gène *RBBP6* code pour une protéine appartenant à la famille des « suppresseurs de tumeur » [191].

Notre étude nous a donc permis d'obtenir des résultats ouvrant de nouvelles perspectives sur la compréhension des mécanismes régissant la sévérité des lentigines sur le visage, et à plus large échelle du vieillissement cutané.

1.2 Travaux portant sur l'affaissement de la paupière

1.2.1 Rappels des résultats

Nous avons aussi réalisé une étude « génome entier » se focalisant sur l'affaissement de la paupière supérieure. Cette étude nous a permis de mettre en évidence plusieurs associations significatives entre des SNPs situés sur les chromosomes 10 et 15 et la sévérité de l'affaissement de la paupière supérieure.

Les SNPs significatifs sur le chromosome 10, rs4746957 ($P = 1{,}69 \times 10^{-8}$) et rs16927253 ($P = 2{,}23 \times 10^{-10}$), sont localisés dans des parties introniques du gène *H2AFY2* et à proximité des gènes *AIFM2, TYSND1,* et *SAR1A*. Les individus homozygotes pour l'allèle mineur de ces SNPs présentaient le grade d'affaissement le plus sévère suggérant un effet récessif de l'allèle mineur. L'analyse des données imputées a permis l'identification de 10 SNPs et d'un indel significativement associés à la sévérité de l'affaissement de la paupière supérieure. Tous ces polymorphismes, en fort déséquilibre de liaison avec les SNPs rs4746957 et rs16927253 ($r^2 > 0{,}8$), sont également situés dans une partie intronique du gène *H2AFY2*.

Une association significative a été obtenue avec le SNP rs4842897 ($P = 9{,}13 \times 10^{-12}$). Ce SNP est localisé dans une région intergénique sur le chromosome 15 à proximité directe des gènes *KLHL25* (< 10 kb) et *AKAP13* (< 54kb) et du microARN *MIR-1276* (< 33 kb). La distribution des génotypes du SNP rs4842897 en fonction de la sévérité de l'affaissement de la paupière a mis en évidence un effet récessif de l'allèle mineur. Deux SNPs imputés, en fort déséquilibre de liaison avec rs4842897 ($r^2 > 0{,}8$), avaient une p-valeur d'association significative. Ces SNPs imputés sont également situés dans une région intergénique du chromosome 15.

1.2.2 Comparaison avec la littérature

Récemment, une GWAS a étudié les facteurs intrinsèques et extrinsèques liés à l'affaissement de la paupière [167]. Cette étude, menée sur une cohorte de 2186 jumeaux anglais et une cohorte de 5578 néerlandais, a mis en évidence le rôle protecteur de l'allèle C du SNP rs11876749, situé à proximité du gène *TGIF1*. Ce SNP n'est pas répliqué dans notre cohorte (P = 0.43). Malheureusement, dans cette publication, la liste des p-valeurs d'association n'est pas renseignée rendant ainsi impossible la réplication de nos résultats avec cette étude.

Cette étude récente est très intéressante car les auteurs ont clairement montré à partir des données génétiques des jumeaux que l'affaissement de la paupière avait une composante génétique très forte (pourcentage d'héritabilité de 61%). Néanmoins, les associations publiées dans cette GWAS, si elles ont bien été obtenues sur la cohorte de néerlandais, ne sont pas répliquées sur la cohorte des jumeaux anglais. De plus, alors que cette étude a montré l'impact de plusieurs facteurs environnementaux sur l'affaissement de la paupière, ceux-ci n'ont apparemment pas été inclus comme co-variables dans les modèles de régression utilisés pour mesurer les associations.

1.2.3 Interprétation biologique

La mise en évidence d'une association significative entre l'affaissement de la paupière supérieure et des SNPs dans le gène *H2AFY2*, codant pour la protéine macroH2A2, une histone de la famille H2A, représente un résultat tout à fait intéressant à plusieurs titres. En effet, plusieurs études ont déjà pointé le rôle des histones et de l'épigénétique dans la sénescence et le vieillissement cellulaire et le vieillissement général [118, 192, 193]. De plus, le gène *H2AFY,* exprimé dans la peau et associé au développement du lupus érythémateux, présente également des profils d'expression différents chez les cellules jeunes et les cellules âgées [194]. Cependant, le rôle des macroH2A2 dans les mécanismes contrôlant l'affaissement de la paupière, et plus généralement le vieillissement, n'est pas encore clairement établi à l'heure actuelle.

Le gène *AIFM2* code pour un facteur d'induction de l'apoptose [195], le gène *TYSND1* contrôle la bêta-oxydation des acides gras via la synthèse d'une protéase [196] et le gène *SAR1A* est impliqué dans le trafic vésiculaire [197].

Dans cette même région génétique, d'autres SNPs présentent des p-valeurs proches du seuil de significativité et sont localisés à proximité du gène *COL13A1* qui code pour la chaine alpha

du collagène de type XIII et contrôlant le maintien des tissus conjonctifs. Le gène *COL13A1* est exprimé chez l'Homme non seulement dans la peau et dans les fibroblastes mais aussi dans l'œil et plus particulièrement dans les muscles squelettiques ciliaires et différents nerfs [198]. Deux études menées chez la souris ont permis de mieux comprendre les rôles du collagène de type XIII, en démontrant notamment son implication dans l'adhésion entre les fibres musculaires et la membrane basale [199] et dans la maturation et le maintien des fonctions neuromusculaires squelettiques [200]. Le gène *COL1A31* serait un bon candidat pour être impliqué dans l'affaissement de la paupière selon deux mécanismes hypothétiques. Le premier consisterait en une déficience ou un changement du taux d'expression du collagène de type XIII pouvant conduire à un manque d'adhésion du muscle à la membrane basale et donc à une accumulation de graisses dans le tissu conjonctif de la paupière responsable de son affaissement. Le second mécanisme reposerait sur une modification de la jonction synaptique neuromusculaire qui pourrait engendrer une déficience du muscle releveur de la paupière.

Les SNPs significatifs localisés sur le chromosome 15 sont situés à proximité directe des gènes *KLHL25* (< 10 kb) et *AKAP13* (< 54kb) et du microARN *MIR-1276* (< 33 kb).

Il est intéressant de noter que le gène *KLHL25*, codant pour la protéine ENCP2 (Ectoderm-Neural Cortex Protein 2), est exprimé dans les tissus conjonctifs et en particulier les tissus adipeux. Même si les connaissances sur la famille de gènes « Kelch-like » et en particulier sur la protéine ENCP2 sont très limitées [201], celle-ci a déjà été associée à l'âge biologique mesuré par un système de score osséographique [180].

Le gène *AKAP13* permet la synthèse de protéines de la famille AKAP (A-Kinase Anchor Protein) qui se lient à des sous-unités régulatrices des kinases et favorisent l'activité dépendante des ligands des récepteurs aux œstrogènes [202]. Ceci représente un résultat intéressant puisque diverses études ont montré qu'une diminution du taux d'œstrogènes accélère le vieillissement cutané, en induisant une baisse du taux de collagène cutané [131].

Finalement, notre recherche bibliographique n'a pas permis de révéler d'information biologique pertinente pour le microARN *MIR-1276* microRNA.

1.3 Recherche des voies de signalisation associées aux différents indicateurs de vieillissement cutané

1.3.1 Rappels des résultats principaux

Parmi les 139 voies de signalisation identifiées, 45 voies de signalisation ont été associées avec au moins un des 4 indicateurs étudiés : Larnier, lentigines, rides, relâchement cutané. Nous avons trouvé des associations significatives (FDR < 25%) avec :

- 18 voies de signalisation pour le grade de photo-vieillissement cutané ;
- 17 voies de signalisation pour le score de relâchement cutané ;
- 10 voies de signalisation pour les rides ;
- 17 voies de signalisation pour le score de lentigines.

Si on regarde les q-valeurs les plus fortes, seuls les indicateurs de Larnier et celui des rides montrent des q-valeurs inférieures à 0,05 avec 5 voies présentant un FDR < 5% dont une voie commune aux deux : la réparation de l'ADN par excision de bases.

Parmi toutes les voies identifiées, certaines présentent une réelle pertinence biologique.

Pour le grade de photo-vieillissement, on pourra citer notamment les voies Wnt (FDR = 4,8%) et de la mélanogenèse (FDR = 12,1%). Ces deux voies portent une pertinence biologique car Wnt intervient dans la mélanogenèse et il est établi qu'une trop forte exposition solaire, principal facteur environnemental responsable du photo-vieillissement, stimule la production de mélanine.

Pour le relâchement cutané, l'association significative obtenue pour la voie de signalisation mTOR (FDR = 8,0%) constitue un résultat intéressant puisque celle-ci a déjà été liée au vieillissement global. De même, l'association de voies liées aux lipides mérite des investigations plus approfondies afin de déterminer leur rôle potentiel dans le relâchement de la peau.

Concernant le score de rides, on pourra citer les voies de signalisation relatives au métabolisme des lipides ainsi que des voies de signalisation relatives au traitement de l'information génétique.

Pour les lentigines, on pourra citer notamment les voies de la mélanogenèse (FDR = 22,6%) et plusieurs voies de signalisation du système immunitaire. Ces résultats sont cohérents avec les connaissances existantes sur les mécanismes responsables de la formation et de la sévérité des lentigines.

1.3.2 Interprétation systémique des résultats précédents

Les voies de signalisations de la base de données KEGG sont divisées en 6 grandes classes: « Traitement de l'information génétique » (« Genetic Information Processing »), « Métabolisme » (« Metabolism »), « Maladies » (« Diseases »), « Traitement de l'information environnementale » (« Environmental Information Processing »), « Processus cellulaires » (« Cellular Processes ») et « Systèmes de l'organisme » (« Organismal Systems »).

Pour avoir une vision d'ensemble, nous avons essayé de voir l'impact relatif de chacune des classes sur chaque indicateur. Pour cela, l'ensemble des voies passant le FDR < 25% pour chaque indicateur a été distribué selon les 6 classes, en pourcentage sur le nombre de voies (Figure 37).

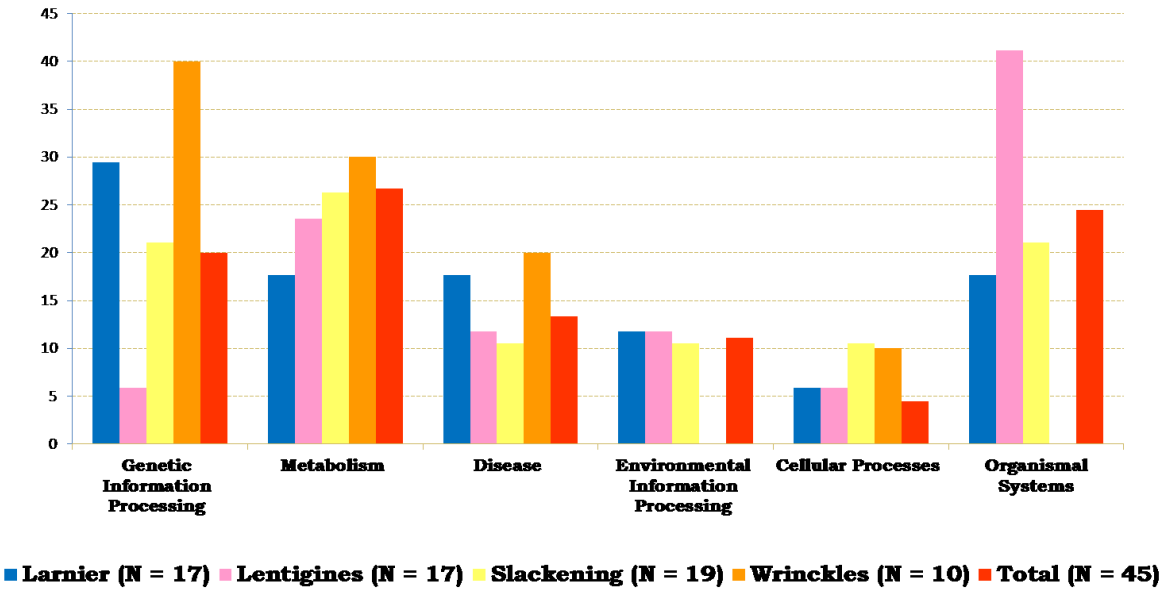


Figure 37. Distribution selon les différentes catégories de voies de signalisation des résultats significativement associées aux différents phénotypes : photo-vieillissement cutané en bleu (Larnier), lentigines en rose, relâchement en jaune (slackening) et rides en orange (wrinckles). La liste de voies de signalisation unique est représentée en rouge (Total).

Tout d'abord, à l'exception des lentigines (≈ 5%), plus de 20% des voies de signalisation associées significativement aux phénotypes étudiés appartiennent à la catégorie « Traitement de l'information génétique ». Cette classe regroupe des voies de signalisation intervenant dans les processus de transcription, de traduction, de repliement, transport et dégradation des protéines et des ARNs, et de réplication et réparation de l'ADN. Ceci représente un résultat intéressant mais aussi cohérent par rapport aux mécanismes connus du vieillissement cutané. En effet, les rayons UV, principal facteur environnemental du vieillissement cutané, engendrent des dommages au niveau de l'ADN dans les cellules exposées [124-126]. De plus, la déficience des mécanismes de réparation de l'ADN ont d'ores et déjà été associés à une accélération du vieillissement [147].

Pour chacun des quatre phénotypes, entre 17 et 30% selon l'indicateur considéré des voies de signalisations associées sont liées au métabolisme. La catégorie « Métabolisme » inclut des voies de signalisation :

- conduisant au métabolisme de différents composés (glucides, lipides, acides aminés, nucléotides, co-facteurs et vitamines, terpénoïdes et polycétides) ;
- conduisant au métabolisme énergétique ;
- impliquées dans la biosynthèse et le métabolisme des glycanes et de certains métabolites secondaires, des voies de signalisation ;
- associées au métabolisme et à la dégradation des xénobiotiques.

Le métabolisme semble donc jouer un rôle non négligeable dans les mécanismes contrôlant le vieillissement cutané. En particulier, les résultats individuels ont souligné l'importance des voies de signalisation relatives au métabolisme des lipides qui sont un constituant essentiel de l'épiderme. Plusieurs études avaient déjà souligné l'impact du métabolisme, et notamment des lipides, dans le vieillissement cutané [147, 148].

De même, mais dans une moindre mesure, 5 à 10% des voies de signalisations identifiées comme associées à chacun des quatre phénotypes appartiennent à la catégorie « Processus cellulaires ». Les voies de signalisations du transport et du catabolisme, de la motilité et la communication cellulaire et de la croissance et la mort cellulaire sont regroupées dans cette

classe. Cette apparente faible implication de ces voies de signalisation peut paraître étonnante puisque la sénescence cellulaire est en première ligne du vieillissement.

Environ 10% des voies de signalisation associées aux phénotypes, rides mises à part (0%), sont répertoriées comme « Traitement de l'information environnementale ». Cette classe contient les voies de signalisation de transport membranaire, de transduction du signal et des molécules de signalisation et d'interaction. Cette proportion relativement faible peut surprendre dans la mesure où les facteurs environnementaux impactent fortement le vieillissement cutané. En même temps, les facteurs environnementaux ont été mis en co-variables dans nos études, et cela peut induire une sous-estimation de leur impact en termes de voies de signalisation biologiques.

Environ 15%, 20% et 40% des voies de signalisation associées respectivement au photo-vieillissement, au relâchement et aux lentigines, sont des voies de signalisation « Systèmes de l'organisme ». Ces différents systèmes sont les systèmes immunitaire, endocrinien, nerveux, digestif, circulatoire, excrétoire, sensoriel, et les systèmes dirigeant le développement et l'adaptation à l'environnement. Cette proportion est particulièrement élevée pour les lentigines, ce qui semble pertinent dans la mesure où le système immunitaire (voir Résultats, chapitre 1) et les mécanismes liés à l'inflammation interviendrait dans l'apparition de ces tâches pigmentaires [203]. Tout comme pour la catégorie « Traitement de l'information environnementale », aucune association entre les rides et les voies de signalisation « Systèmes de l'organisme » n'a été obtenue.

Au total, ces résultats montrent clairement qu'il existe un socle commun de mécanismes intervenant dans le vieillissement cutané mais qu'il existe aussi des spécificités propres à chacun des indicateurs du vieillissement cutané. Ceci est particulièrement visible pour les rides puisqu'aucun résultat significatif n'est obtenu avec les classes « Traitement de l'information environnementale » et « Systèmes de l'organisme » contrairement aux 3 autres phénotypes, alors qu'une proportion plus importante de voies de signalisation « Traitement de l'information génétique » est observée. De la même façon, la proportion de voies de signalisation « Systèmes de l'organisme » associées aux lentigines est largement supérieure à celles obtenues avec les autres phénotypes mais en contrepartie, il semble que les processus de « Traitement de l'information génétique » soient moins impliqués que dans les autres phénotypes.

Nos résultats nous ont permis d'identifier des voies de signalisation biologiques associées avec nos indicateurs du vieillissement cutané. Certaines d'entre elles sont spécifiques de chacun des indicateurs : 7 pour le photo-vieillissement cutané, 10 pour les lentigines, 12 pour le relâchement et 3 pour les rides. D'autres, au contraire, sont communes à 2 ou 3 des indicateurs. Aucune voie de signalisation n'est commune aux 4 indicateurs (Figure 38).

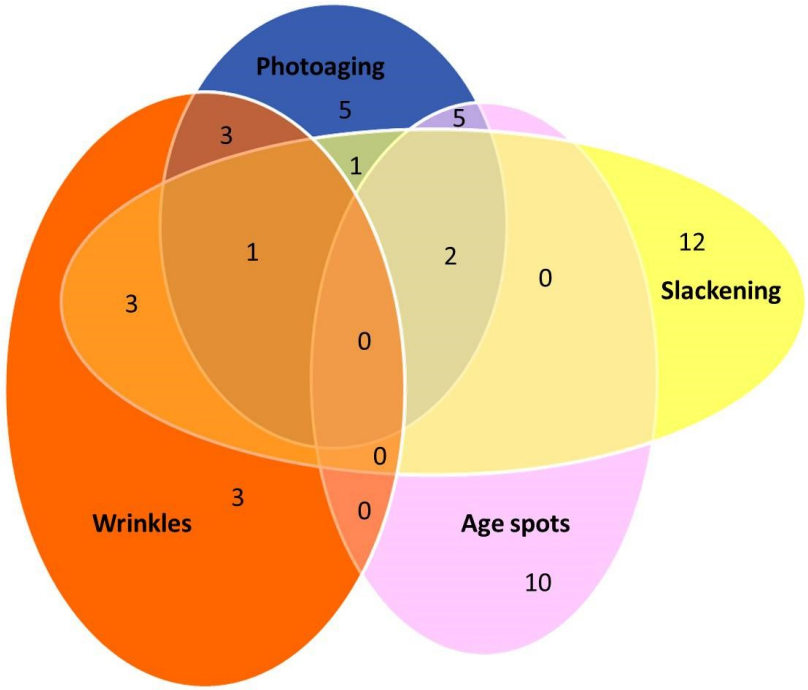


Figure 38. Nombre de voies de signalisation communes et exclusives aux différents phénotypes.

Nous nous sommes intéressés plus particulièrement aux voies de signalisation communes à 3 indicateurs. Deux voies de signalisation ont été identifiées comme étant significativement associées avec le photo-vieillissement, les lentigines et le relâchement. Il s'agit des voies de signalisation de la mélanogenèse et de la régulation du cytosquelette d'actine.

La mélanogenèse est le processus de création de la mélanine, responsable de la pigmentation. Ce résultat est pertinent puisque la mélanogenèse est un des principaux mécanismes de défense contre les rayons UV, eux-mêmes principaux responsables du vieillissement cutané extrinsèque.

De même, l'association avec la régulation du cytosquelette d'actine semble pertinente. Les filaments d'actine sont un constituant essentiel du cytosquelette des cellules eucaryotes ainsi que des fibres musculaires. Le cytosquelette est l'ensemble des polymères qui organise la

structure de la cellule et lui confère une partie de ses propriétés mécaniques. L'exposition aux rayons UV, accélérant le vieillissement cutané, a pour conséquence la perturbation de la structure du cytosquelette, et notamment des filaments d'actine [204].

La voie de signalisation de la réparation par excision de nucléotide, qui permet la réparation d'ADN endommagé, est associée avec le photo-vieillissement, les rides et le relâchement. Une fois de plus, ce résultat est cohérent avec le phénomène étudié. D'une part, les mécanismes de réparation, et en particulier celui de l'ADN, perdent de leur efficacité avec le passage du temps et, d'autre part, les rayons UV entrainent des erreurs dans l'ADN.

Ces observations nécessitent des analyses supplémentaires plus fines (fonction des polymorphismes, impact sur l'expression…) afin de mieux appréhender les mécanismes aboutissant au vieillissement cutané. Enfin, il est important de souligner que de telles études de voies de signalisation biologique n'ont, à notre connaissance, pas encore été publiées dans le cadre du vieillissement cutané. Actuellement, nous sommes en train d'essayer de répliquer ces résultats avec d'autres méthodes (voir Discussion, chapitre 3.3).

2 Critique des méthodes utilisées

2.1 Critique des études « génome-entier »

L'avènement des études « génome-entier » démontre leur grande utilité même si ce type d'analyse présente quelques inconvénients.

2.1.1 Intérêt

Le grand avantage des études « génome-entier » réside dans le fait que celles-ci permettent une recherche exhaustive des variants associés à un caractère, en excluant tout a priori biologique. Par conséquent, il a été possible d'associer des régions chromosomiques à certains caractères alors que leur lien était jusque-là insoupçonné.

Un deuxième avantage de ces études est leur relative facilité d'exécution. En effet, les méthodologies utilisées lors d'une GWAS sont désormais bien établies et le protocole à suivre pour réaliser une telle étude est bien défini. En effet, plusieurs publications détaillent précisément les différentes étapes à suivre lors d'une GWAS, aussi bien au niveau du contrôle qualité qu'au niveau de l'analyse statistique [55, 205-207].

Enfin, le développement des techniques de biologie moléculaire et de bioinformatique a largement contribué à l'essor des études « génome-entier ». En effet, les puces de génotypage permettent désormais de génotyper jusqu'à 2,5 millions de SNPs chez plusieurs milliers d'individus à moindre coût. Par conséquent, les GWAS sont désormais des analyses relativement peu coûteuses permettant de recueillir une grande quantité d'information. En parallèle, les améliorations de la bioinformatique, et notamment les méthodes d'imputation, ont rendu plus aisée la comparaison des résultats obtenus sur différentes cohortes et ont également facilité la mise en place de méta-analyse.

Néanmoins, les études « génome-entier » sont encore confrontées à quelques inconvénients, notamment d'un point de vue statistique.

2.1.2 Limites

L'inconvénient principal des études « génome-entier » apparaît lors de la correction des tests multiples permettant de repérer les polymorphismes significativement associés au phénotype

analysé. En effet, la méthode la plus couramment utilisée pour pallier le problème de comparaison multiple est la méthode de Bonferroni (voir Introduction, chapitre 5.3.2.2). Bien qu'elle soit simple et rapide à mettre en place, cette méthode est très (trop) conservative. Seuls les signaux très associés seront détectés, conférant ainsi une confiance certaine en les résultats. Néanmoins, des signaux en réalité associés au caractère étudié mais dont les effets sont plus modestes peuvent ne pas être considérés comme significativement associés à cause de la stringence de cette correction. L'approche alternative du taux de fausse découverte permet d'atténuer ce phénomène, mais seulement de manière partielle. Avoir à disposition des outils statistiques permettant une meilleure correction des tests multiples présenterait alors l'avantage de ne pas laisser de côté des signaux réels mais plus modestes.

Une autre limite de ces méthodes découle directement du paradigme selon lequel les maladies fréquentes sont les conséquences de polymorphismes fréquents. Ainsi, les puces de génotypage ont été élaborées de façon à privilégier les SNPs dont la MAF est supérieure à 5%. Les SNPs moins fréquents sont de fait exclus de ces analyses. Or, depuis quelques années, ce paradigme est remis en question puisque les variants communs ne suffisent pas à déterminer complètement l'architecture génétique des pathologies complexes [208]. L'imputation a en partie rendu possible l'analyse des SNPs de fréquence faible, en permettant l'inférence des génotypes de ces SNPs mais la qualité de leur imputation peut s'avérer problématique. Se pose également le problème de la puissance statistique lors de l'analyse de tels polymorphismes. En effet, un grand nombre de patients est nécessaire afin d'établir la significativité statistique d'une association entre un SNP de fréquence faible et un phénotype. Là encore, une solution pourrait être la mise en commun des données issues de différentes cohortes dans le but d'entreprendre des méta-analyses.

2.2 Critique des analyses des voies de signalisation

Les analyses visant à rechercher des voies de signalisation associées à un phénotype présentent des avantages permettant de répondre aux limites des GWAS mais ces méthodes, relativement récentes et toujours en cours de développement, ne sont pas encore tout à fait au point.

2.2.1 Intérêt

L'avantage principal des recherches de voies de signalisation associées à un phénotype réside dans l'intégration d'a priori biologique par l'intermédiaire de listes de gènes impliquées dans

une même fonction. L'interprétation biologique des résultats obtenus est alors plus aisée. En effet, au contraire d'un simple polymorphisme dont l'impact biologique est parfois difficile à saisir, les associations significatives identifiées correspondent directement à une fonction biologique définie. En conséquence, ces analyses permettent une meilleure compréhension des mécanismes mis en jeu dans le phénotype étudié.

De plus, ce type d'analyse permet de mettre en évidence des listes de gènes associées à un phénotype qui n'auraient pas été repérées par une GWAS classique. Une des forces de ces analyses est donc leur capacité à repérer des polymorphismes aux effets plus faibles et non « détectables » par les GWAS mais dont l'enrichissement au sein d'une même voie biologique permet de les associer au phénotype. En effet, l'action conjointe de SNPs dont les effets individuels sont faibles peut jouer un rôle significatif sur le caractère étudié. Alors que les études « génome-entier » ne permettent pas la détection de tels SNPs, les analyses des voies biologiques semblent particulièrement adaptées à ce type de situation.

Enfin, ce type d'analyse est moins soumis aux problématiques liées à la correction des tests multiples. Etant donné que ces études se focalisent sur des listes de gènes et non sur des polymorphismes ponctuels, le nombre de comparaisons effectuées est moindre et ce type d'analyse est donc moins conservative.

Néanmoins, les recherches d'association de voies de signalisation présentent plusieurs limites.

2.2.2 Limites

Les limites de ces méthodologies sont d'une part d'ordre biologique et d'autre part statistique.

Dans un premier temps, la connaissance sur les voies de signalisation n'est que partielle. Bien que quelques voies de signalisation soient bien documentées, un grand nombre de réseaux biologiques restent encore inconnus ou mal renseignés. Cela entraîne une perte d'information, puisqu'un grand nombre de SNPs étudiés dans les GWAS ne peut être associé à aucune voie de signalisation. En outre, les voies de signalisation sont principalement établies à partir des relations existantes entre les gènes dans les phénomènes de transcription ou d'interactions au niveau des protéines. D'autres facteurs de régulation, comme les facteurs épigénétiques ou les modifications post-transcriptionnelles par exemple, ne sont pas toujours pris en compte dans l'élaboration des réseaux biologiques. Enfin, un certain nombre de voies de signalisation sont établies à partir de prédictions informatiques dont la fiabilité est moindre comparée à celle des

réseaux obtenus à partir de preuve expérimentale. La qualité des réseaux biologiques est donc inégale, et certaines voies de signalisation étudiées peuvent être imprécises.

En plus de ces limites biologiques, plusieurs contraintes statistiques sont à prendre en considération. Plusieurs sources de biais existent au cours des différentes étapes de ces analyses. Par exemple, les gènes contenant un grand nombre de SNPs ont a priori plus de chance d'inclure des SNPs significatifs par le simple fait du hasard. Ce problème se pose également pour les voies de signalisation composées d'un grand nombre de SNPs. Des stratégies de permutations sont utilisées pour s'affranchir de cette problématique mais elles possèdent elles aussi leurs limites et peuvent également créer des biais dans l'analyse [93]. De plus, les méthodes visant à assigner une p-valeur à un gène peuvent être source de biais. Par exemple, dans l'algorithme GSEA, la p-valeur d'association assignée à un gène est la p-valeur du SNP le plus fortement associé au phénotype situé dans le gène. Cela peut-être problématique puisque dans le cas où cette p-valeur est le fruit du hasard, le gène sera considéré comme associé à tort au phénotype. D'autres méthodes, et notamment des tests prenant en compte l'ensemble des polymorphismes situés dans le gène, ont été mises en place afin d'attribuer une p-valeur à chaque gène. Cependant, elles ne sont pas implémentées dans toutes les méthodologies et peuvent être délicates à mettre en place.

Plusieurs revues énumèrent les principales limites des techniques développées pour identifier des voies de signalisation associées à un phénotype [93-97].

3 Perspectives

3.1 Analyse des CNVs

Les puces de génotypage donnent également des informations concernant les Copy Number Variations. Ainsi, les données génomiques relatives au vieillissement cutané dont nous disposons comportent les génotypes de 91 706 marqueurs permettant la détection des CNVs. Ces polymorphismes ont été exclus des analyses présentées dans ce travail puisque leur inférence à l'aide des puces de génotypage est complexe. Cependant, leur analyse ne doit pas être omise dans la mesure où des variants de cette nature ont déjà été associés à plusieurs maladies comme le cancer [195, 196], l'autisme [197] ou l'obésité [193], et que leur influence sur la variabilité observée face aux maladies est reconnue. Bien que l'étude des CNVs soit encore problématique à partir des données obtenues avec les puces, il semble essentiel de fournir ce travail sur les données génomiques du vieillissement cutané.

3.2 Différence entre l'âge chronologique et l'âge perçu par la peau sur le visage

En plus des 12 indicateurs de vieillissement cutané renseignés dans les données analysées, l'âge des patientes a été estimé à partir de la peau du visage. Ainsi, la différence entre l'âge chronologique réel de la patiente et son âge estimé fournit une information supplémentaire concernant le vieillissement. Récemment, une GWAS a analysé la différence entre ces deux mesures de l'âge et a permis l'identification de plusieurs gènes promouvant la jeunesse cutanée [169]. Mener des études semblables à celles réalisées au cours de ma thèse permettra alors de répliquer les signaux obtenus par la GWAS déjà publiée et peut-être d'identifier de nouvelles régions chromosomiques impactant le vieillissement cutané du visage.

3.3 Réplication des résultats des voies de signalisation

Jusqu'à présent, la recherche de voies de signalisation et de réseaux biologiques associés aux différents indicateurs de vieillissement cutané a été effectuée à l'aide la méthode GSEA. Cependant, il est primordial de répliquer ces premiers résultats à l'aide d'autres méthodes afin de s'assurer de leur fiabilité. De plus, l'utilisation de méthodes différentes permet de combler les limites de l'algorithme GSEA et peut ainsi rendre possible l'obtention de nouvelles

associations. Actuellement, la méthode SNP Ratio Test (SRT) qui compare la proportion de SNPs déclarés significatifs au sein de chaque voie de signalisation est appliquée à nos données afin de répliquer les résultats déjà obtenus. Enfin, il est également prévu d'utiliser des méthodes plus récentes, multivariées ou basées sur la topologie des voies de signalisation [209], afin de compléter ce travail.

3.4 Réplications et méta-analyse

La réplication des résultats obtenus lors d'une GWAS est essentielle. Elle consiste à comparer les p-valeurs des signaux obtenus à celles obtenues sur une cohorte indépendante étudiant le même phénotype. Les méta-analyses combinent les données génomiques de différentes études afin d'augmenter la puissance statistique et permet de confirmer les signaux obtenus individuellement dans chaque étude et d'obtenir de nouveaux signaux. Plusieurs GWAS sur le vieillissement cutané ont été publiées ces 3 dernières années [167, 169], et il serait donc intéressant de développer des collaborations entre les différents partenaires à l'origine de ces études afin de mettre en application ces stratégies.

3.5 Dynamique du vieillissement cutané

Les données analysées ont permis l'identification de régions chromosomiques associées à la sévérité du vieillissement cutané. Récemment, les femmes de la cohorte SU.VI.MAX ayant accepté de participer à cette étude ont été de nouveau contactées afin de réévaluer les indicateurs de vieillissement cutané 10 ans après la première mesure. La comparaison de ces deux mesures rend alors possible la définition de plusieurs estimations de la vitesse du vieillissement de la peau sur le visage. Si le nombre de femmes pour lesquelles les nouvelles mesures sont obtenues est suffisamment important pour assurer une certaine puissance statistique, cela permettrait d'identifier des régions chromosomiques potentiellement associées à une accélération, ou un ralentissement, du vieillissement cutané.

3.6 Approches multi-marqueurs

Les études d'association « génome-entier » analysent des marqueurs simples. Or, pour des maladies complexes, la combinaison de plusieurs polymorphismes, et non un seul, peut être à l'origine de la maladie. Il semble donc pertinent de mettre en place des stratégies permettant d'évaluer ces possibilités.

3.6.1 Haplotypes

Les haplotypes sont définis comme la combinaison de polymorphismes situés sur un même chromosome (voir Introduction, chapitre 2.3). Ces haplotypes peuvent être à l'origine de dysfonctionnement et ainsi expliquer certaines maladies [210-213]. Ces exemples illustrent la pertinence de la recherche d'haplotypes comme facteur de risque de certaines pathologies. Toutefois, de nombreuses questions sont sous-jacentes à ces analyses. En effet, comment définir les haplotypes étudiés ? Combien et quels marqueurs inclure dans les haplotypes considérés ? Enfin, étant donné le nombre d'haplotypes qu'il est possible d'analyser, comment contourner les limites des corrections multiples afin de ne pas être trop conservatif ?

3.6.2 Epistasie

L'épistasie désigne l'interaction entre deux loci : l'impact d'un génotype à un locus particulier dépend du génotype à un autre locus. Les GWAS ont pour vocation de rechercher les polymorphismes ayant individuellement un impact sur le caractère étudié mais ceux-ci ne représentent qu'une fraction des facteurs génétiques responsables du phénotype [208]. Les interactions entre SNPs peuvent expliquer une autre portion de ces facteurs génétiques. Par exemple, des interactions entre le gène *PTPN2* et des gènes du métabolisme de la vitamine D contribuent au risque d'arthrite chronique juvénile [214]. De même, une interaction entre le gène *ERAP1* et le gène *HLA-C* influence la susceptibilité au psoriasis [215]. Ces deux exemples illustrent la pertinence de la recherche d'épistasies pour déterminer les facteurs génétiques impliqués dans des maladies complexes.

Plusieurs méthodes ont été implémentées afin de rechercher les épistasies [178, 216-220]. Cependant, la recherche d'interactions est encore confrontée à des problèmes statistiques tels que la puissance ou la correction des tests multiples. Une publication récente détaille les principales méthodes existantes et les problèmes inhérents à ces techniques [221].

3.7 Nouvelles technologies de séquençage

Les avancées passées de biologie moléculaire et de bioinformatique ont permis le développement des études d'association « génome-entier ». Aujourd'hui, l'émergence de nouvelles technologies a rendu possible le séquençage intégral du génome ou de ses parties codantes rapidement chez un grand nombre de personnes comparativement aux technologies plus anciennes. Ces nouvelles méthodes permettent la caractérisation des polymorphismes de fréquence faible et ont ouvert de nouvelles perspectives dans de nombreuses maladies [199].

Par exemple, le séquençage intégral dans le cadre d'une étude de liaison a permis de répliquer les variants causaux du syndrome de Miller et ainsi valider cette nouvelle approche [201]. De même, le séquençage des exons du génome a abouti à l'identification de plusieurs variants impliqués dans différentes maladies [200, 222-224].

Ces nouvelles technologies permettent la détection de variants de fréquence faible ou rares ayant des effets forts sur le phénotype étudié et complètent ainsi les GWAS qui se focalisent principalement sur les variants communs. Toutefois, de nombreux défis accompagnent ces technologies, que ce soit du point de vue informatique, bioinformatique, statistique ou encore financier. En effet, la quantité de données générées par ces séquençages est considérable et nécessite des capacités de stockage et une structure informatique adaptée. L'analyse de ces données requiert également des améliorations d'un point de vue bioinformatique afin d'une part de réduire les temps de calcul et d'autre part d'implémenter les méthodes adaptées. De plus, ces technologies sont particulièrement efficaces lorsqu'elles sont employées dans des analyses de liaison mais des problèmes quant à la puissance statistique se posent lorsqu'elles sont utilisées dans des études d'association en raison de la faible fréquence des variants et nécessitent alors le développement de méthodes plus adaptées [225]. Enfin, l'aspect financier ne doit pas être négligé. En effet, bien que les coûts de ces technologies aient sensiblement diminué ces dernières années, le séquençage « génome-entier » reste encore bien plus onéreux que le génotypage à l'aide d'une puce.

3.8 Biologie des systèmes

Les études d'association « génome-entier » permettent la détection de polymorphismes génétiques associés à un phénotype. Toutefois, l'intégration de données biologiques est une étape indispensable afin de mieux appréhender les mécanismes moléculaires sous-jacents. Plusieurs disciplines ont vocation à générer des données permettant de répondre à ces interrogations, et notamment :

- L'épigénomique étudie les mécanismes de régulation de l'expression des gènes pouvant être influencés par l'environnement ;
- Le transcriptome étudie les ARNs messagers issus de l'expression d'une partie du génome. La comparaison des données transcriptomiques et des données génomiques a notamment permis l'élaboration de bases de données renseignant l'impact de polymorphismes sur l'expression des gènes [84, 85, 87] ;

- Le protéome étudie l'ensemble des protéines présentes dans la cellule ;
- Le métabolome étudie de petites molécules comme les hormones par exemple dans différents tissus.

Le projet ENCODE [226, 227], qui a débuté en 2003, a pour ambition de décrire les éléments fonctionnels du génome humain en combinant ces différentes disciplines. Les données d'ores et déjà générées sont publiques et ont été exploitées dans différentes bases de données comme RegulomeDB [228]. L'émergence de la biologie systémique ouvre de nombreuses perspectives pour une meilleure compréhension des mécanismes biologiques.

Conclusion

Depuis mon arrivée au sein du laboratoire GBA en 2010, j'ai pu assister à l'essor des analyses de données génomiques. L'amélioration des techniques a permis le génotypage de millions de polymorphismes et il est maintenant possible de séquencer rapidement le génome de plusieurs milliers d'individus. Tous ces progrès de biologie moléculaire s'accompagnent de leurs défis aussi bien d'un point de vue statistique qu'informatique, faisant ainsi de la bioinformatique une discipline en constante évolution.

Les travaux réalisés dans le cadre de cette thèse illustrent ce phénomène. D'une part, j'ai appliqué des méthodologies clairement établies par l'intermédiaire d'une étude d'association « génome entier ». D'autre part, j'ai recherché des associations avec les voies de signalisation qui sont plus novatrices et pour lesquelles aucun protocole défini n'a encore émergé au sein de la communauté scientifique.

Une première étude d'association « génome-entier » a permis d'associer le gène *HLA-C* à la sévérité des lentigines sur le visage. Ce résultat suggère ainsi un rôle du système immunitaire dans ce phénotype en réponse notamment aux dommages engendrés par les rayons UV. Une seconde GWAS a permis l'identification d'une association significative entre le gène *H2AFY2* codant pour une histone de la famille H2A et l'affaissement de la paupière.

La recherche de voies de signalisation associées aux différents indicateurs de vieillissement sur le visage en est encore à un stade préliminaire mais la pertinence des résultats obtenus à l'aide de l'algorithme GSEA confirme l'intérêt de telles analyses. Néanmoins, il est indispensable de refaire ces analyses en utilisant d'autres méthodes afin de confirmer les associations obtenues et peut-être en identifier de nouvelles.

L'identification des mécanismes moléculaires du vieillissement cutané présente beaucoup d'intérêt. D'une part, cela peut permettre de suggérer de nouvelles approches cosmétiques pour prévenir le vieillissement cutané. D'autre part, comme la peau est un bon modèle du vieillissement général, ces résultats peuvent ouvrir de nouvelles perspectives sur la compréhension du vieillissement global qui constitue un axe majeur de la recherche actuelle.

Mon travail de thèse m'a permis d'évoluer dans un domaine à l'interface de la biologie, la génétique, la statistique et la bioinformatique. J'ai eu la chance de pouvoir mettre en œuvre plusieurs méthodologies dans un domaine en perpétuelle évolution et dans lequel beaucoup reste encore à faire. Au total, ces années de thèse m'ont donné l'opportunité de compléter ma formation initiale au sein d'un laboratoire dynamique à l'image de la bioinformatique.

Bibliographie

[1]. Chargaff, E. *Chemical specificity of nucleic acids and mechanism of their enzymatic degradation.* Experientia, 1950. **6**(6): p. 201-9.

[2]. Chargaff, E. *Some recent studies on the composition and structure of nucleic acids.* J Cell Physiol Suppl, 1951. **38**(Suppl. 1): p. 41-59.

[3]. Watson, J.D. and Crick, F.H. *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.

[4]. Abecasis, G.R., Altshuler, D., Auton, A., et al. *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

[5]. Sherry, S.T., Ward, M.H., Kholodov, M., et al. *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.

[6]. Redon, R., Ishikawa, S., Fitch, K.R., et al. *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-54.

[7]. Hardy, G.H. *Mendelian proportions in a mixed population.* Science, 1908. **28**(706): p. 49-50.

[8]. Weinberg, W. *Über den Nachweis der Vererbung beim Menschen.* Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg, 1908. **64**.

[9]. Robbins, R.B. *Some Applications of Mathematics to Breeding Problems III.* Genetics, 1918. **3**(4): p. 375-89.

[10]. Lewontin, R.C. *The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.* Genetics, 1964. **49**(1): p. 49-67.

[11]. Hill, W.G. and Robertson, A. *Linkage disequilibrium in finite populations.* Theor Appl Genet, 1968. **38**(6): p. 226-31.

[12]. The International HapMap Consortium. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

[13]. Myers, S., Bottolo, L., Freeman, C., et al. *A fine-scale map of recombination rates and hotspots across the human genome.* Science, 2005. **310**(5746): p. 321-4.

[14]. Myers, S., Spencer, C.C., Auton, A., et al. *The distribution and causes of meiotic recombination in the human genome.* Biochem Soc Trans, 2006. **34**(Pt 4): p. 526-30.

[15]. Gabriel, S.B., Schaffner, S.F., Nguyen, H., et al. *The structure of haplotype blocks in the human genome.* Science, 2002. **296**(5576): p. 2225-9.

[16]. Zhang, W., Hui, K.Y., Gusev, A., et al. *Extended haplotype association study in Crohn's disease identifies a novel, Ashkenazi Jewish-specific missense mutation in the NF-kappaB pathway gene, HEATR3.* Genes Immun, 2013. **14**(5): p. 310-6.

[17]. Clark, A.G. *The role of haplotypes in candidate gene studies.* Genet Epidemiol, 2004. **27**(4): p. 321-33.

[18]. Khoury, M.J., Little, J., and Burke, W. *Human Genome Epidemiology: Scope and Strategies*, in *Human Genome Epidemiology: A Scientific Foundation for Using Genetic to Improve Health and Prevent Disease*, M.J. Khoury, J. Little, and W. Burke, Editors. 2003, Oxford University Press. p. 3-16.

[19]. Kerem, B., Rommens, J.M., Buchanan, J.A., et al. *Identification of the cystic fibrosis gene: genetic analysis.* Science, 1989. **245**(4922): p. 1073-80.

[20]. The Huntington's Disease Collaborative Research Group. *A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes.* Cell, 1993. **72**(6): p. 971-83.

[21]. Yu, J.T., Tan, L., and Hardy, J. *Apolipoprotein E in Alzheimer's disease: an update.* Annu Rev Neurosci, 2014. **37**: p. 79-100.

[22]. Bordeleau, L., Panchal, S., and Goodwin, P. *Prognosis of BRCA-associated breast cancer: a summary of evidence.* Breast Cancer Res Treat, 2010. **119**(1): p. 13-24.

[23]. Limou, S., Le Clerc, S., Coulonges, C., et al. *Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02).* J Infect Dis, 2009. **199**(3): p. 419-26.

[24]. Balding, D.J. *A tutorial on statistical methods for population association studies.* Nat Rev Genet, 2006. **7**(10): p. 781-91.

[25]. Riordan, J.R., Rommens, J.M., Kerem, B., et al. *Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA.* Science, 1989. **245**(4922): p. 1066-73.

[26]. Rommens, J.M., Iannuzzi, M.C., Kerem, B., et al. *Identification of the cystic fibrosis gene: chromosome walking and jumping.* Science, 1989. **245**(4922): p. 1059-65.

[27]. Bertram, L. and Tanzi, R.E. *The genetic epidemiology of neurodegenerative disease.* J Clin Invest, 2005. **115**(6): p. 1449-57.

[28]. Nakamura, S. *[Huntington's disease--advances in gene mapping]*. Nihon Rinsho, 1993. **51**(9): p. 2481-7.

[29]. Gusella, J.F., Wexler, N.S., Conneally, P.M., et al. *A polymorphic DNA marker genetically linked to Huntington's disease.* Nature, 1983. **306**(5940): p. 234-8.

[30]. Mira, M.T., Alcais, A., di Pietrantonio, T., et al. *Segregation of HLA/TNF region is linked to leprosy clinical spectrum in families displaying mixed leprosy subtypes.* Genes Immun, 2003. **4**(1): p. 67-73.

[31]. Simon-Sanchez, J., Schulte, C., Bras, J.M., et al. *Genome-wide association study reveals genetic risk underlying Parkinson's disease.* Nat Genet, 2009. **41**(12): p. 1308-12.

[32]. Soto-Ortolaza, A.I., Heckman, M.G., Labbe, C., et al. *GWAS risk factors in Parkinson's disease: LRRK2 coding variation and genetic interaction with PARK16.* Am J Neurodegener Dis, 2013. **2**(4): p. 287-99.

[33]. Easton, D.F. and Eeles, R.A. *Genome-wide association studies in cancer.* Hum Mol Genet, 2008. **17**(R2): p. R109-15.

[34]. Song, H., Koessler, T., Ahmed, S., et al. *Association study of prostate cancer susceptibility variants with risks of invasive ovarian, breast, and colorectal cancer.* Cancer Res, 2008. **68**(21): p. 8837-42.

[35]. Le Clerc, S., Limou, S., Coulonges, C., et al. *Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03).* J Infect Dis, 2009. **200**(8): p. 1194-201.

[36]. Patnala, R., Clements, J., and Batra, J. *Candidate gene association studies: a comprehensive guide to useful in silico tools.* BMC Genet, 2013. **14**: p. 39.

[37]. Dean, M., Carrington, M., Winkler, C., et al. *Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study.* Science, 1996. **273**(5283): p. 1856-62.

[38]. Samson, M., Libert, F., Doranz, B.J., et al. *Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene.* Nature, 1996. **382**(6593): p. 722-5.

[39]. Seghatoleslam, A., Bozorg-Ghalati, F., Monabati, A., et al. *UBE2Q1, as a Down Regulated Gene in Pediatric Acute Lymphoblastic Leukemia.* Int J Mol Cell Med, 2014. **3**(2): p. 95-101.

[40]. Welter, D., MacArthur, J., Morales, J., et al. *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.

[41]. van Manen, D., Delaneau, O., Kootstra, N.A., et al. *Genome-wide association scan in HIV-1-infected individuals identifying variants influencing disease course.* PLoS One, 2011. **6**(7): p. e22208.

[42]. Nakayama, E.E., Meyer, L., Iwamoto, A., et al. *Protective effect of interleukin-4 -589T polymorphism on human immunodeficiency virus type 1 disease progression: relationship with virus load.* J Infect Dis, 2002. **185**(8): p. 1183-6.

[43]. Kwa, D., van Rij, R.P., Boeser-Nunnink, B., et al. *Association between an interleukin-4 promoter polymorphism and the acquisition of CXCR4 using HIV-1 variants.* AIDS, 2003. **17**(7): p. 981-5.

[44]. Namiki, T., Tanemura, A., Valencia, J.C., et al. *AMP kinase-related kinase NUAK2 affects tumor growth, migration, and clinical outcome of human melanoma.* Proc Natl Acad Sci U S A, 2011. **108**(16): p. 6597-602.

[45]. Wu, X., Ye, Y., Rosell, R., et al. *Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy.* J Natl Cancer Inst, 2011. **103**(10): p. 817-25.

[46]. Xun, W.W., Brennan, P., Tjonneland, A., et al. *Single-nucleotide polymorphisms (5p15.33, 15q25.1, 6p22.1, 6q27 and 7p15.3) and lung cancer survival in the European Prospective Investigation into Cancer and Nutrition (EPIC).* Mutagenesis, 2011. **26**(5): p. 657-66.

[47]. Morgan, T.M., House, J.A., Cresci, S., et al. *Investigation of 95 variants identified in a genome-wide study for association with mortality after acute coronary syndrome.* BMC Med Genet, 2011. **12**: p. 127.

[48]. Ripatti, S., Tikkanen, E., Orho-Melander, M., et al. *A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses.* Lancet, 2010. **376**(9750): p. 1393-400.

[49]. Collet, B. *Modelling Survival Data in Medical Research, Second Edition*. 2003: Chapman and Hall/CRC. 409.

[50]. Klein, R.J., Zeiss, C., Chew, E.Y., et al. *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.

[51]. The International HapMap Consortium. *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.

[52]. The International HapMap Consortium. *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.

[53]. The International HapMap Consortium. *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.

[54]. Manolio, T.A. and Collins, F.S. *The HapMap and genome-wide association studies in diagnosis and therapy.* Annu Rev Med, 2009. **60**: p. 443-56.

[55]. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., et al. *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.

[56]. Wang, L. and Xu, Y. *Haplotype inference by maximum parsimony.* Bioinformatics, 2003. **19**(14): p. 1773-80.

[57]. Bafna, V., Gusfield, D., Lancia, G., et al. *Haplotyping as perfect phylogeny: a direct approach.* J Comput Biol, 2003. **10**(3-4): p. 323-40.

[58]. Halperin, E. and Eskin, E. *Haplotype reconstruction from genotype data using Imperfect Phylogeny.* Bioinformatics, 2004. **20**(12): p. 1842-9.

[59]. Niu, T., Qin, Z.S., Xu, X., et al. *Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms.* Am J Hum Genet, 2002. **70**(1): p. 157-69.

[60]. Zhao, J.H. and Sham, P.C. *Faster haplotype frequency estimation using unrelated subjects.* Hum Hered, 2002. **53**(1): p. 36-41.

[61]. Stephens, M., Smith, N.J., and Donnelly, P. *A new statistical method for haplotype reconstruction from population data.* Am J Hum Genet, 2001. **68**(4): p. 978-89.

[62]. Delaneau, O., Zagury, J.F., and Marchini, J. *Improved whole-chromosome phasing for disease and population genetic studies.* Nat Methods, 2013. **10**(1): p. 5-6.

[63]. Delaneau, O. and Zagury, J.F. *Haplotype inference.* Methods Mol Biol, 2012. **888**: p. 177-96.

[64]. Adkins, R.M. *Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset.* BMC Genet, 2004. **5**: p. 22.

[65]. Browning, S.R. and Browning, B.L. *Haplotype phasing: existing methods and new developments.* Nat Rev Genet, 2011. **12**(10): p. 703-14.

[66]. Salem, R.M., Wessel, J., and Schork, N.J. *A comprehensive literature review of haplotyping software and methods for use with unrelated individuals.* Hum Genomics, 2005. **2**(1): p. 39-66.

[67]. Niu, T. *Algorithms for inferring haplotypes.* Genet Epidemiol, 2004. **27**(4): p. 334-47.

[68]. Howie, B.N., Donnelly, P., and Marchini, J. *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.

[69]. Howie, B., Marchini, J., and Stephens, M. *Genotype imputation with thousands of genomes.* G3 (Bethesda), 2011. **1**(6): p. 457-70.

[70]. Wigginton, J.E., Cutler, D.J., and Abecasis, G.R. *A note on exact tests of Hardy-Weinberg equilibrium.* Am J Hum Genet, 2005. **76**(5): p. 887-93.

[71]. Pritchard, J.K., Stephens, M., and Donnelly, P. *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.

[72]. Price, A.L., Patterson, N.J., Plenge, R.M., et al. *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.

[73]. Bland, J.M. and Altman, D.G. *Multiple significance tests: the Bonferroni method.* BMJ, 1995. **310**(6973): p. 170.

[74]. Benjamini, Y. and Hochberg, Y. *Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.

[75]. Storey, J.D. and Tibshirani, R. *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.

[76]. Forner, K., Lamarine, M., Guedj, M., et al. *Universal false discovery rate estimation methodology for genome-wide association studies.* Hum Hered, 2008. **65**(4): p. 183-94.

[77]. Guedj, M., Robelin, D., Hoebeke, M., et al. *Detecting local high-scoring segments: a first-stage approach for genome-wide association studies.* Stat Appl Genet Mol Biol, 2006. **5**: p. Article22.

[78]. Dalmasso, C., Bar-Hen, A., and Broet, P. *A constrained polynomial regression procedure for estimating the local False Discovery Rate.* BMC Bioinformatics, 2007. **8**: p. 229.

[79]. Baecklund, F., Foo, J.N., Bracci, P., et al. *A comprehensive evaluation of the role of genetic variation in follicular lymphoma survival.* BMC Med Genet, 2014. **15**(1): p. 113.

[80]. Troyer, J.L., Nelson, G.W., Lautenberger, J.A., et al. *Genome-wide association study implicates PARD3B-based AIDS restriction.* J Infect Dis, 2011. **203**(10): p. 1491-502.

[81]. Benitez-Parejo, N., Rodriguez del Aguila, M.M., and Perez-Vicente, S. *Survival analysis and Cox regression.* Allergol Immunopathol (Madr), 2011. **39**(6): p. 362-73.

[82]. Berger, V.W., Stefanescu, C., and Zhou, Y.Y. *The analysis of stratified 2 x 2 contingency tables.* Biom J, 2006. **48**(6): p. 992-1007.

[83]. Clayton, D. *Link functions in multi-locus genetic models: implications for testing, prediction, and interpretation.* Genet Epidemiol, 2012. **36**(4): p. 409-18.

[84]. Dimas, A.S., Deutsch, S., Stranger, B.E., et al. *Common regulatory variation impacts gene expression in a cell type-dependent manner.* Science, 2009. **325**(5945): p. 1246-50.

[85]. Dixon, A.L., Liang, L., Moffatt, M.F., et al. *A genome-wide association study of global gene expression.* Nat Genet, 2007. **39**(10): p. 1202-7.

[86]. Grundberg, E., Small, K.S., Hedman, A.K., et al. *Mapping cis- and trans-regulatory effects across multiple tissues in twins.* Nat Genet, 2012. **44**(10): p. 1084-9.

[87]. Zeller, T., Wild, P., Szymczak, S., et al. *Genetics and beyond--the transcriptome of human monocytes and disease susceptibility.* PLoS One, 2010. **5**(5): p. e10693.

[88]. McLaren, P.J., Coulonges, C., Ripke, S., et al. *Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls.* PLoS Pathog, 2013. **9**(7): p. e1003515.

[89]. Cerhan, J.R., Berndt, S.I., Vijai, J., et al. *Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma.* Nat Genet, 2014.

[90]. Springelkamp, H., Hohn, R., Mishra, A., et al. *Meta-analysis of genome-wide association studies identifies novel loci that influence cupping and the glaucomatous process.* Nat Commun, 2014. **5**: p. 4883.

[91]. Al Olama, A.A., Kote-Jarai, Z., Berndt, S.I., et al. *A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer.* Nat Genet, 2014. **46**(10): p. 1103-9.

[92]. Hysi, P.G., Cheng, C.Y., Springelkamp, H., et al. *Genome-wide analysis of multi-ancestry cohorts identifies new loci influencing intraocular pressure and susceptibility to glaucoma.* Nat Genet, 2014. **46**(10): p. 1126-30.

[93]. Wang, K., Li, M., and Hakonarson, H. *Analysing biological pathways in genome-wide association studies.* Nat Rev Genet, 2010. **11**(12): p. 843-54.

[94]. Ramanan, V.K., Shen, L., Moore, J.H., et al. *Pathway analysis of genomic data: concepts, methods, and prospects for future development.* Trends Genet, 2012. **28**(7): p. 323-32.

[95]. Fridley, B.L. and Biernacka, J.M. *Gene set analysis of SNP data: benefits, challenges, and future directions.* Eur J Hum Genet, 2011. **19**(8): p. 837-43.

[96]. Evangelou, M., Rendon, A., Ouwehand, W.H., et al. *Comparison of methods for competitive tests of pathway analysis.* PLoS One, 2012. **7**(7): p. e41018.

[97]. Khatri, P., Sirota, M., and Butte, A.J. *Ten years of pathway analysis: current approaches and outstanding challenges.* PLoS Comput Biol, 2012. **8**(2): p. e1002375.

[98]. Jumpertz, R., Hanson, R.L., Sievers, M.L., et al. *Higher energy expenditure in humans predicts natural mortality.* J Clin Endocrinol Metab, 2011. **96**(6): p. E972-6.

[99]. Folsom, A.R., Kaye, S.A., Sellers, T.A., et al. *Body fat distribution and 5-year risk of death in older women.* JAMA, 1993. **269**(4): p. 483-7.

[100]. Kohrt, W.M., Kirwan, J.P., Staten, M.A., et al. *Insulin resistance in aging is related to abdominal obesity.* Diabetes, 1993. **42**(2): p. 273-81.

[101]. Evans, W.J., Paolisso, G., Abbatecola, A.M., et al. *Frailty and muscle metabolism dysregulation in the elderly.* Biogerontology, 2010. **11**(5): p. 527-36.

[102]. Khan, A.S., Sane, D.C., Wannenburg, T., et al. *Growth hormone, insulin-like growth factor-1 and the aging cardiovascular system.* Cardiovasc Res, 2002. **54**(1): p. 25-35.

[103]. Sonntag, W.E., Lynch, C., Thornton, P., et al. *The effects of growth hormone and IGF-1 deficiency on cerebrovascular and brain ageing.* J Anat, 2000. **197 Pt 4**: p. 575-85.

[104]. Hayflick, L. *The limited in vitro lifetime of humain diploid cell strains.* Exp Cell Res, 1965. **37**: p. 614-36.

[105]. Campisi, J. *Senescent cells, tumor suppression, and organismal aging: good citizens, bad neighbors.* Cell, 2005. **120**(4): p. 513-22.

[106]. Trougakos, I.P., Saridaki, A., Panayotou, G., et al. *Identification of differentially expressed proteins in senescent human embryonic fibroblasts.* Mech Ageing Dev, 2006. **127**(1): p. 88-92.

[107]. Shelton, D.N., Chang, E., Whittier, P.S., et al. *Microarray analysis of replicative senescence.* Curr Biol, 1999. **9**(17): p. 939-45.

[108]. Campisi, J. and d'Adda di Fagagna, F. *Cellular senescence: when bad things happen to good cells.* Nat Rev Mol Cell Biol, 2007. **8**(9): p. 729-40.

[109]. Lopez-Otin, C., Blasco, M.A., Partridge, L., et al. *The hallmarks of aging.* Cell, 2013. **153**(6): p. 1194-217.

[110]. Koga, H., Kaushik, S., and Cuervo, A.M. *Protein homeostasis and aging: The importance of exquisite quality control.* Ageing Res Rev, 2011. **10**(2): p. 205-15.

[111]. Cadenas, E. and Davies, K.J. *Mitochondrial free radical generation, oxidative stress, and aging.* Free Radic Biol Med, 2000. **29**(3-4): p. 222-30.

[112]. Barja, G. *The mitochondrial free radical theory of aging.* Prog Mol Biol Transl Sci, 2014. **127**: p. 1-27.

[113]. Lauri, A., Pompilio, G., and Capogrossi, M.C. *The mitochondrial genome in aging and senescence.* Ageing Res Rev, 2014. **18C**: p. 1-15.

[114]. Burtner, C.R. and Kennedy, B.K. *Progeria syndromes and ageing: what is the connection?* Nat Rev Mol Cell Biol, 2010. **11**(8): p. 567-78.

[115]. Hoeijmakers, J.H. *DNA damage, aging, and cancer.* N Engl J Med, 2009. **361**(15): p. 1475-85.

[116]. Worman, H.J. *Nuclear lamins and laminopathies.* J Pathol, 2012. **226**(2): p. 316-25.

[117]. Armanios, M. and Blackburn, E.H. *The telomere syndromes.* Nat Rev Genet, 2012. **13**(10): p. 693-704.

[118]. Das, C. and Tyler, J.K. *Histone exchange and histone modifications during transcription and aging.* Biochim Biophys Acta, 2013. **1819**(3-4): p. 332-42.

[119]. Dimauro, T. and David, G. *Chromatin modifications: the driving force of senescence and aging?* Aging (Albany NY), 2009. **1**(2): p. 182-90.

[120]. Farage, M.A., Miller, K.W., and Maibach, H.I. *Degenerative Changes in Aging Skin*, in *Textbook of Aging Skin*, M.A. Farage, K.W. Miller, and H.I. Maibach, Editors. 2010, Springer Berlin Heidelberg. p. 25-35.

[121]. Raschke, C. and Elsner, P. *Skin Aging: A Brief Summary of Characteristic Changes*, in *Textbook of Aging Skin*, M.A. Farage, K.W. Miller, and H.I. Maibach, Editors. 2010, Springer Berlin Heidelberg. p. 37-43.

[122]. Jenkins, G. *Molecular mechanisms of skin ageing.* Mech Ageing Dev, 2002. **123**(7): p. 801-10.

[123]. Farage, M.A., Miller, K.W., Elsner, P., et al. *Intrinsic and extrinsic factors in skin ageing: a review.* Int J Cosmet Sci, 2008. **30**(2): p. 87-95.

[124]. Rabe, J.H., Mamelak, A.J., McElgunn, P.J., et al. *Photoaging: mechanisms and repair.* J Am Acad Dermatol, 2006. **55**(1): p. 1-19.

[125]. Yaar, M. and Gilchrest, B.A. *Photoageing: mechanism, prevention and therapy.* Br J Dermatol, 2007. **157**(5): p. 874-87.

[126]. Sage, E., Girard, P.M., and Francesconi, S. *Unravelling UVA-induced mutagenesis.* Photochem Photobiol Sci, 2012. **11**(1): p. 74-80.

[127]. Guyuron, B., Rowe, D.J., Weinfeld, A.B., et al. *Factors contributing to the facial aging of identical twins.* Plast Reconstr Surg, 2009. **123**(4): p. 1321-31.

[128]. Kennedy, C., Bastiaens, M.T., Bajdik, C.D., et al. *Effect of smoking and sun on the aging skin.* J Invest Dermatol, 2003. **120**(4): p. 548-54.

[129]. Purba, M.B., Kouris-Blazos, A., Wattanapenpaiboon, N., et al. *Skin wrinkling: can food make a difference?* J Am Coll Nutr, 2001. **20**(1): p. 71-80.

[130]. Farage, M.A., Miller, K.W., Elsner, P., et al. *Characteristics of the Aging Skin.* Adv Wound Care (New Rochelle), 2013. **2**(1): p. 5-10.

[131]. Shah, M.G. and Maibach, H.I. *Estrogen and skin. An overview.* Am J Clin Dermatol, 2001. **2**(3): p. 143-50.

[132]. Stevenson, S. and Thornton, J. *Effect of estrogens on skin aging and the potential role of SERMs.* Clin Interv Aging, 2007. **2**(3): p. 283-97.

[133]. Ashcroft, G.S. and Ashworth, J.J. *Potential role of estrogens in wound healing.* Am J Clin Dermatol, 2003. **4**(11): p. 737-43.

[134]. Thornton, M.J. *The biological actions of estrogens on skin.* Exp Dermatol, 2002. **11**(6): p. 487-502.

[135]. Thornton, M.J. *Oestrogen functions in skin and skin appendages.* Expert Opin Ther Targets, 2005. **9**(3): p. 617-29.

[136]. Thornton, M.J. *Estrogens and aging skin.* Dermatoendocrinol, 2013. **5**(2): p. 264-70.

[137]. Brincat, M., Versi, E., Moniz, C.F., et al. *Skin collagen changes in postmenopausal women receiving different regimens of estrogen therapy.* Obstet Gynecol, 1987. **70**(1): p. 123-7.

[138]. Affinito, P., Palomba, S., Sorrentino, C., et al. *Effects of postmenopausal hypoestrogenism on skin collagen.* Maturitas, 1999. **33**(3): p. 239-47.

[139]. Vaillant, L. and Callens, A. *[Hormone replacement treatment and skin aging]*. Therapie, 1996. **51**(1): p. 67-70.

[140]. Punnonen, R. *Effect of castration and peroral estrogen therapy on the skin.* Acta Obstet Gynecol Scand Suppl, 1972. **21**: p. 3-44.

[141]. Maheux, R., Naud, F., Rioux, M., et al. *A randomized, double-blind, placebo-controlled study on the effect of conjugated estrogens on skin thickness.* Am J Obstet Gynecol, 1994. **170**(2): p. 642-9.

[142]. Makrantonaki, E., Schonknecht, P., Hossini, A.M., et al. *Skin and brain age together: The role of hormones in the ageing process.* Exp Gerontol, 2010. **45**(10): p. 801-13.

[143]. Goldstein, S., Moerman, E.J., Jones, R.A., et al. *Insulin-like growth factor binding protein 3 accumulates to high levels in culture medium of senescent and quiescent human fibroblasts.* Proc Natl Acad Sci U S A, 1991. **88**(21): p. 9680-4.

[144]. Yoon, I.K., Kim, H.K., Kim, Y.K., et al. *Exploration of replicative senescence-associated genes in human dermal fibroblasts by cDNA microarray technology.* Exp Gerontol, 2004. **39**(9): p. 1369-78.

[145]. Shekar, S.N., Luciano, M., Duffy, D.L., et al. *Genetic and environmental influences on skin pattern deterioration.* J Invest Dermatol, 2005. **125**(6): p. 1119-29.

[146]. Makrantonaki, E., Adjaye, J., Herwig, R., et al. *Age-specific hormonal decline is accompanied by transcriptional changes in human sebocytes in vitro.* Aging Cell, 2006. **5**(4): p. 331-44.

[147]. Lener, T., Moll, P.R., Rinnerthaler, M., et al. *Expression profiling of aging in the human skin.* Exp Gerontol, 2006. **41**(4): p. 387-97.

[148]. Zouboulis, C.C. and Makrantonaki, E. *Clinical aspects and molecular diagnostics of skin aging.* Clin Dermatol, 2011. **29**(1): p. 3-14.

[149]. Makrantonaki, E., Brink, T.C., Zampeli, V., et al. *Identification of biomarkers of human skin ageing in both genders. Wnt signalling - a label of skin ageing?* PLoS One, 2012. **7**(11): p. e50393.

[150]. Carlson, M.E., Silva, H.S., and Conboy, I.M. *Aging of signal transduction pathways, and pathology.* Exp Cell Res, 2008. **314**(9): p. 1951-61.

[151]. DeCarolis, N.A., Wharton, K.A., Jr., and Eisch, A.J. *Which way does the Wnt blow? Exploring the duality of canonical Wnt signaling on cellular aging.* Bioessays, 2008. **30**(2): p. 102-6.

[152]. Elfakir, A., Ezzedine, K., Latreille, J., et al. *Functional MC1R-gene variants are associated with increased risk for severe photoaging of facial skin.* J Invest Dermatol, 2010. **130**(4): p. 1107-15.

[153]. Suppa, M., Elliott, F., Mikeljevic, J.S., et al. *The determinants of periorbital skin ageing in participants of a melanoma case-control study in the U.K.* Br J Dermatol, 2011. **165**(5): p. 1011-21.

[154]. Box, N.F., Wyeth, J.R., O'Gorman, L.E., et al. *Characterization of melanocyte stimulating hormone receptor variant alleles in twins with red hair.* Hum Mol Genet, 1997. **6**(11): p. 1891-7.

[155]. Valverde, P., Healy, E., Jackson, I., et al. *Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans.* Nat Genet, 1995. **11**(3): p. 328-30.

[156]. Smith, R., Healy, E., Siddiqui, S., et al. *Melanocortin 1 receptor variants in an Irish population.* J Invest Dermatol, 1998. **111**(1): p. 119-22.

[157]. Flanagan, N., Healy, E., Ray, A., et al. *Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation.* Hum Mol Genet, 2000. **9**(17): p. 2531-7.

[158]. Bastiaens, M., ter Huurne, J., Gruis, N., et al. *The melanocortin-1-receptor gene is the major freckle gene.* Hum Mol Genet, 2001. **10**(16): p. 1701-8.

[159]. Duffy, D.L., Box, N.F., Chen, W., et al. *Interactive effects of MC1R and OCA2 on melanoma risk phenotypes.* Hum Mol Genet, 2004. **13**(4): p. 447-61.

[160]. Motokawa, T., Kato, T., Hashimoto, Y., et al. *Effect of Val92Met and Arg163Gln variants of the MC1R gene on freckles and solar lentigines in Japanese.* Pigment Cell Res, 2007. **20**(2): p. 140-3.

[161]. Valverde, P., Healy, E., Sikkink, S., et al. *The Asp84Glu variant of the melanocortin 1 receptor (MC1R) is associated with melanoma.* Hum Mol Genet, 1996. **5**(10): p. 1663-6.

[162]. Palmer, J.S., Duffy, D.L., Box, N.F., et al. *Melanocortin-1 receptor polymorphisms and risk of melanoma: is the association explained solely by pigmentation phenotype?* Am J Hum Genet, 2000. **66**(1): p. 176-86.

[163]. Kennedy, C., ter Huurne, J., Berkhout, M., et al. *Melanocortin 1 receptor (MC1R) gene variants are associated with an increased risk for cutaneous melanoma which is largely independent of skin type and hair color.* J Invest Dermatol, 2001. **117**(2): p. 294-300.

[164]. Matichard, E., Verpillat, P., Meziani, R., et al. *Melanocortin 1 receptor (MC1R) gene variants may increase the risk of melanoma in France independently of clinical risk factors and UV exposure.* J Med Genet, 2004. **41**(2): p. e13.

[165]. Stratigos, A.J., Dimisianos, G., Nikolaou, V., et al. *Melanocortin receptor-1 gene polymorphisms and the risk of cutaneous melanoma in a low-risk southern European population.* J Invest Dermatol, 2006. **126**(8): p. 1842-9.

[166]. Le Clerc, S., Taing, L., Ezzedine, K., et al. *A genome-wide association study in Caucasian women points out a putative role of the STXBP5L gene in facial photoaging.* J Invest Dermatol, 2013. **133**(4): p. 929-35.

[167]. Jacobs, L.C., Liu, F., Bleyen, I., et al. *Intrinsic and Extrinsic Risk Factors for Sagging Eyelids.* JAMA Dermatol, 2014.

[168]. Rittie, L. and Fisher, G.J. *UV-light-induced signal cascades and skin aging.* Ageing Res Rev, 2002. **1**(4): p. 705-20.

[169]. Chang, A.L., Atzmon, G., Bergman, A., et al. *Identification of genes promoting skin youthfulness by genome-wide association study.* J Invest Dermatol, 2014. **134**(3): p. 651-7.

[170]. Bione, S., Sala, C., Manzini, C., et al. *A human homologue of the Drosophila melanogaster diaphanous gene is disrupted in a patient with premature ovarian failure: evidence for conserved function in oogenesis and implications for human sterility.* Am J Hum Genet, 1998. **62**(3): p. 533-41.

[171]. Marozzi, A., Manfredini, E., Tibiletti, M.G., et al. *Molecular definition of Xq common-deleted region in patients affected by premature ovarian failure.* Hum Genet, 2000. **107**(4): p. 304-11.

[172]. Liu, Y.L., Lu, W.C., Brummel, T.J., et al. *Reduced expression of alpha-1,2-mannosidase I extends lifespan in Drosophila melanogaster and Caenorhabditis elegans.* Aging Cell, 2009. **8**(4): p. 370-9.

[173]. Hercberg, S., Galan, P., Preziosi, P., et al. *Background and rationale behind the SU.VI.MAX Study, a prevention trial using nutritional doses of a combination of antioxidant vitamins and minerals to reduce cardiovascular diseases and cancers. SUpplementation en VItamines et Mineraux AntioXydants Study.* Int J Vitam Nutr Res, 1998. **68**(1): p. 3-20.

[174]. Hercberg, S. *The SU.VI.MAX study, a randomized, placebo-controlled trial on the effects of antioxidant vitamins and minerals on health.* Ann Pharm Fr, 2006. **64**(6): p. 397-401.

[175]. Larnier, C., Ortonne, J.P., Venot, A., et al. *Evaluation of cutaneous photodamage using a photographic scale.* Br J Dermatol, 1994. **130**(2): p. 167-73.

[176]. Morizot F, L.S., Guinot C, Binder M, et al. *Development of photographic scales documenting features of skin ageing based on digital images.* Ann Dermatol Venereol, 2002, 129(Suppl 1 Part 2) :1s402.

[177]. Guinot C, Malvy D, Latreille J et al. *Sun exposure behaviour of a general adult population in France*. In: Ring J, Weidinger S, Darsow U eds Skin and Environment - Perception and Protection. Monduzzi editore S.p.A: Bologna, 2001. 1099–106

[178]. Purcell, S., Neale, B., Todd-Brown, K., et al. *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

[179]. Marchini, J., Howie, B., Myers, S., et al. *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.

[180]. Lunetta, K.L., D'Agostino, R.B., Sr., Karasik, D., et al. *Genetic correlates of longevity and selected age-related phenotypes: a genome-wide association study in the Framingham Study.* BMC Med Genet, 2007. **8 Suppl 1**: p. S13.

[181]. Jia, X., Han, B., Onengut-Gumuscu, S., et al. *Imputing amino acid polymorphisms in human leukocyte antigens.* PLoS One, 2013. **8**(6): p. e64683.

[182]. Kanehisa, M. and Goto, S. *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

[183]. Wang, K., Li, M., and Bucan, M. *Pathway-based approaches for analysis of genomewide association studies.* Am J Hum Genet, 2007. **81**(6): p. 1278-83.

[184]. Ezzedine, K., Mauger, E., Latreille, J., et al. *Freckles and solar lentigines have different risk factors in Caucasian women.* J Eur Acad Dermatol Venereol, 2013. **27**(3): p. e345-56.

[185]. Johnson, S.C., Rabinovitch, P.S., and Kaeberlein, M. *mTOR is a key modulator of ageing and age-related disease.* Nature, 2013. **493**(7432): p. 338-45.

[186]. Zhang, G., Li, J., Purkayastha, S., et al. *Hypothalamic programming of systemic ageing involving IKK-beta, NF-kappaB and GnRH.* Nature, 2013. **497**(7448): p. 211-6.

[187]. Lin, C.B., Babiarz, L., Liebel, F., et al. *Modulation of microphthalmia-associated transcription factor gene expression alters skin pigmentation.* J Invest Dermatol, 2002. **119**(6): p. 1330-40.

[188]. Hirobe, T. *How are proliferation and differentiation of melanocytes regulated?* Pigment Cell Melanoma Res, 2011. **24**(3): p. 462-78.

[189]. Kvedar, J.C., Manabe, M., Phillips, S.B., et al. *Characterization of sciellin, a precursor to the cornified envelope of human keratinocytes.* Differentiation, 1992. **49**(3): p. 195-204.

[190]. Akerstrom, B., Logdberg, L., Berggard, T., et al. *alpha(1)-Microglobulin: a yellow-brown lipocalin.* Biochim Biophys Acta, 2000. **1482**(1-2): p. 172-84.

[191]. Motadi, L.R., Bhoola, K.D., and Dlamini, Z. *Expression and function of retinoblastoma binding protein 6 (RBBP6) in human lung cancer.* Immunobiology, 2011. **216**(10): p. 1065-73.

[192]. Oberdoerffer, P. *An age of fewer histones.* Nat Cell Biol, 2010. **12**(11): p. 1029-31.

[193]. Rai, T.S. and Adams, P.D. *Lessons from senescence: Chromatin maintenance in non-proliferating cells.* Biochim Biophys Acta, 2012. **1819**(3-4): p. 322-31.

[194]. Sporn, J.C., Kustatscher, G., Hothorn, T., et al. *Histone macroH2A isoforms predict the risk of lung cancer recurrence.* Oncogene, 2009. **28**(38): p. 3423-8.

[195]. Gong, M., Hay, S., Marshall, K.R., et al. *DNA binding suppresses human AIF-M2 activity and provides a connection between redox chemistry, reactive oxygen species, and apoptosis.* J Biol Chem, 2007. **282**(41): p. 30331-40.

[196]. Kurochkin, I.V., Mizuno, Y., Konagaya, A., et al. *Novel peroxisomal protease Tysnd1 processes PTS1- and PTS2-containing enzymes involved in beta-oxidation of fatty acids.* EMBO J, 2007. **26**(3): p. 835-45.

[197]. Long, K.R., Yamamoto, Y., Baker, A.L., et al. *Sar1 assembly regulates membrane constriction and ER export.* J Cell Biol, 2010. **190**(1): p. 115-28.

[198]. Sandberg-Lall, M., Hagg, P.O., Wahlstrom, I., et al. *Type XIII collagen is widely expressed in the adult and developing human eye and accentuated in the ciliary muscle, the optic nerve and the neural retina.* Exp Eye Res, 2000. **70**(4): p. 401-10.

[199]. Kvist, A.P., Latvanlehto, A., Sund, M., et al. *Lack of cytosolic and transmembrane domains of type XIII collagen results in progressive myopathy.* Am J Pathol, 2001. **159**(4): p. 1581-92.

[200]. Latvanlehto, A., Fox, M.A., Sormunen, R., et al. *Muscle-derived collagen XIII regulates maturation of the skeletal neuromuscular junction.* J Neurosci, 2010. **30**(37): p. 12230-41.

[201]. Dhanoa, B.S., Cogliati, T., Satish, A.G., et al. *Update on the Kelch-like (KLHL) gene family.* Hum Genomics, 2013. **7**(1): p. 13.

[202]. Driggers, P.H., Segars, J.H., and Rubino, D.M. *The proto-oncoprotein Brx activates estrogen receptor beta by a p38 mitogen-activated protein kinase pathway.* J Biol Chem, 2001. **276**(50): p. 46792-7.

[203]. Aoki, H., Moro, O., Tagami, H., et al. *Gene expression profiling analysis of solar lentigo in relation to immunohistochemical characteristics.* Br J Dermatol, 2007. **156**(6): p. 1214-23.

[204]. Provost, N., Moreau, M., Leturque, A., et al. *Ultraviolet A radiation transiently disrupts gap junctional communication in human keratinocytes.* Am J Physiol Cell Physiol, 2003. **284**(1): p. C51-9.

[205]. Anderson, C.A., Pettersson, F.H., Clarke, G.M., et al. *Data quality control in genetic case-control association studies.* Nat Protoc, 2010. **5**(9): p. 1564-73.

[206]. Clarke, G.M., Anderson, C.A., Pettersson, F.H., et al. *Basic statistical analysis in genetic case-control studies.* Nat Protoc, 2011. **6**(2): p. 121-33.

[207]. Bush, W.S. and Moore, J.H. *Chapter 11: Genome-wide association studies.* PLoS Comput Biol, 2012. **8**(12): p. e1002822.

[208]. Maher, B. *Personal genomes: The case of the missing heritability.* Nature, 2008. **456**(7218): p. 18-21.

[209]. Jin, L., Zuo, X.Y., Su, W.Y., et al. *Pathway-based Analysis Tools for Complex Diseases: A Review.* Genomics Proteomics Bioinformatics, 2014. **12**(5): p. 210-220.

[210]. Tregouet, D.A., Konig, I.R., Erdmann, J., et al. *Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease.* Nat Genet, 2009. **41**(3): p. 283-5.

[211]. Wang, Q., Lv, H., Lv, W., et al. *Genome-wide haplotype association study identifies BLM as a risk gene for prostate cancer in Chinese population.* Tumour Biol, 2014.

[212]. Lambert, J.C., Grenier-Boley, B., Harold, D., et al. *Genome-wide haplotype association study identifies the FRMD4A gene as a risk locus for Alzheimer's disease.* Mol Psychiatry, 2013. **18**(4): p. 461-70.

[213]. Hunter, D.J., Kraft, P., Jacobs, K.B., et al. *A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.* Nat Genet, 2007. **39**(7): p. 870-4.

[214]. Ellis, J.A., Scurrah, K.J., Li, Y.R., et al. *Epistasis amongst PTPN2 and genes of the vitamin D pathway contributes to risk of juvenile idiopathic arthritis.* J Steroid Biochem Mol Biol, 2015. **145**: p. 113-20.

[215]. Genetic Analysis of Psoriasis, C., the Wellcome Trust Case Control, C., Strange, A., et al. *A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1.* Nat Genet, 2010. **42**(11): p. 985-90.

[216]. Gyenesei, A., Moody, J., Semple, C.A., et al. *High-throughput analysis of epistasis in genome-wide association studies with BiForce.* Bioinformatics, 2012. **28**(15): p. 1957-64.

[217]. Prabhu, S. and Pe'er, I. *Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease.* Genome Res, 2012. **22**(11): p. 2230-40.

[218]. Zhang, Y. and Liu, J.S. *Bayesian inference of epistatic interactions in case-control studies.* Nat Genet, 2007. **39**(9): p. 1167-73.

[219]. Zhang, X., Huang, S., Zou, F., et al. *TEAM: efficient two-locus epistasis tests in human genome-wide association study.* Bioinformatics, 2010. **26**(12): p. i217-27.

[220]. Herold, C., Steffens, M., Brockschmidt, F.F., et al. *INTERSNP: genome-wide interaction analysis guided by a priori information.* Bioinformatics, 2009. **25**(24): p. 3275-81.

[221]. Wei, W.H., Hemani, G., and Haley, C.S. *Detecting epistasis in human complex traits.* Nat Rev Genet, 2014. **15**(11): p. 722-33.

[222]. Cruchaga, C., Karch, C.M., Jin, S.C., et al. *Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease.* Nature, 2014. **505**(7484): p. 550-4.

[223]. Ng, S.B., Bigham, A.W., Buckingham, K.J., et al. *Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome.* Nat Genet, 2010. **42**(9): p. 790-3.

[224]. Pierce, S.B., Walsh, T., Chisholm, K.M., et al. *Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome.* Am J Hum Genet, 2010. **87**(2): p. 282-8.

[225]. Kiezun, A., Garimella, K., Do, R., et al. *Exome sequencing and the genetic basis of complex traits.* Nat Genet, 2012. **44**(6): p. 623-30.

[226]. ENCODE Project Consortium. *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.

[227]. Qu, H. and Fang, X. *A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project.* Genomics Proteomics Bioinformatics, 2013. **11**(3): p. 135-41.

[228]. Boyle, A.P., Hong, E.L., Hariharan, M., et al. *Annotation of functional variation in personal genomes using RegulomeDB.* Genome Res, 2012. **22**(9): p. 1790-7.

Liste des publications

Laville V.*, Le Clerc S.*, Ezzedine K., Jdid R., Taing L., Labib T., Coulonges T., Ulveling D., Carpentier W., Galan P., Hercberg S., Morizot F., Latreille J., Malvy D., Tschachler E., Guinot C., Zagury J.F. « **A Genome-Wide Association Study in Caucasian Women reveals the involvement of *HLA* genes in the severity of facial solar lentigines** », *Journal of Investigative Dermatology.* - Article en cours de revision.

Laville V.*, Le Clerc S.*, Ezzedine K., Jdid R., Taing L., Labib T., Coulonges T., Ulveling D., Carpentier W., Galan P., Hercberg S., Morizot F., Latreille J., Malvy D., Tschachler E., Guinot C., Zagury J.F. « **A Genome-Wide Association Study identifies new genetic associations in upper eyelid sagging.** » - Article en preparation

Le Clerc S., Delaneau O., Coulonges C., Spadoni J.L., Labib T., **Laville V.**, Ulveling D., Noirel J., Montes M., Schächter F., Caillat-Zucman S., Zagury J.F., « **Evidence After Imputation for a Role of MICA Variants in Nonprogression and Elite Control of HIV Type 1 Infection.** », *Journal Of Infectious Disease*, 2014, 210(12), 1946-50

Lagarde N., Ben Nasr N., Jérémie A., Guillemain H., **Laville V.**, Labib T., Zagury J.F., Montes M. « **NRLiSt BDB, the manually curated nuclear receptors ligands and structure benchmarking database.** », Journal of Medicinal Chemistry, 2014, 57(7), 3117-25.

Liste des communications orales

Laville V. « Exploration and perspectives in the skin ageing genomic project. ». Genomics and disease pathogenesis. Ermenonville, 16-17 décembre 2013.

Laville V. « Genome-wide association studies for skin ageing. ». Workshop: "From Bioinformatics to Therapeutics", Doha, 7-9 avril 2014.

Posters

« A Genome-Wide Association Study in a Caucasian cohort reveals a genetic association between the HLA-C*0701 Allele and Solar Lentigines ». XXII International Pigment Cell Conference, Singapour, 4-7 septembre 2014.

Laville V., Le Clerc S., Ezzedine K., Berlin I., Carpentier W., Jdid R., Galan P., Hercberg S., Guinot C., Morizot F., Latreille J., Tschachler E., Zagury J.F. **« Association génétique entre l'allèle HLA-C*0701 et les lentigines solaires chez une population caucasienne »**. Journées Dermatologiques de Paris, Paris, 9-13 décembre 2014.

Peptinov

Vincent LAVILLE

le cnam

Analyses génomiques de données sur le vieillissement cutané

Résumé

La peau est un excellent modèle d'étude du vieillissement général. En plus de facteurs environnementaux, les facteurs génétiques jouent un rôle majeur dans le vieillissement cutané.

Dans le cadre de ma thèse, j'ai eu accès à une cohorte exceptionnelle de 502 femmes caucasiennes très bien caractérisées sur le plan cutané, pour effectuer deux études d'association « génome-entier ». La première étude a montré le rôle joué par le système immunitaire, et en particulier le gène *HLA-C*, dans la sévérité des lentigines du visage. La seconde a mis en évidence une association entre le gène *H2AFY2* et la sévérité de l'affaissement de la paupière supérieure. La recherche de voies de signalisation biologiques associées à différents indicateurs du vieillissement cutané a souligné le rôle de la mélanogénèse et des mécanismes de réparation de l'ADN.

Ces résultats ouvrent de nouvelles perspectives dans la compréhension des mécanismes inhérents au vieillissement cutané et général.

Mots clés : étude d'association « génome-entier », SNP, voie de signalisation biologique, vieillissement cutané

Résumé en anglais

The skin is an excellent model to study general ageing. In addition to environmental factors, genetic factors play a key role in skin ageing mechanisms.

During my PhD, I have had access to a unique cohort of 502 Caucasian women very-well characterized regarding their facial features to perform two genome-wide association studies. The first one pointed to the role of the immune system, and especially the *HLA-C* gene, in the severity of facial lentigines. The second one identified an association between the *H2AFY2* gene and the severity of superior eyelid drooping. I also looked for associations between biological pathways and several skin ageing indicators which underlined the role of the melanogenesis and several mechanisms of DNA repair.

Overall, these results lead to new insights in the understanding of the molecular mechanisms underlying skin and global ageing.

Keywords: genome-wide association study, GWAS, SNP, biological pathway, skin ageing